Exploiting Big Data in Logistics Risk Assessment via Bayesian Nonparametrics

by

Yan Shang

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
David B. Dunson, Supervisor

_____
Surya T. Tokdar

_____
Jing-Sheng Song

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2014

Abstract

Exploiting Big Data in Logistics Risk Assessment via Bayesian Nonparametrics

by

Yan Shang

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
David B. Dunson, Supervisor

_____
Surya T. Tokdar

_____
Jing-Sheng Song

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistical Science in the Graduate School of Duke University
2014

Abstract

In cargo logistics, a key performance measure is transport risk, defined as the deviation of the actual arrival time from the planned arrival time. Neither earliness nor tardiness is desirable for the customer and freight forwarder. In this paper, we investigate ways to assess and forecast transport risks using a half-year of air cargo data, provided by a leading forwarder on 1336 routes served by 20 airlines. Interestingly, our preliminary data analysis shows a strong multimodal feature in the transport risks, driven by unobserved events, such as cargo missing flights. To accommodate this feature, we introduce a Bayesian nonparametric model – the probit stick-breaking process (PSBP) mixture model – for flexible estimation of the conditional (i.e., state-dependent) density function of transport risk. We demonstrate that using simpler methods, such as OLS linear regression, can lead to misleading inferences. Our model provides a tool for the forwarder to offer customized price and service quotes. It can also generate baseline airline performance to enable fair supplier evaluation. Furthermore, the method allows us to separate recurrent risks from disruption risks. This is important, because hedging strategies for these two kinds of risks are often drastically different.

I dedicate my thesis work to my family and my advisors, Prof. David Dunson and Prof. Jeannette Song, without whom I cannot finish my thesis. I also thanks Prof. Surya Tokdar for his encouragement and support along the way!

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

Here are the distribution used frequently in this study.

$\mathbf{N}(\cdot \mid \mu,\ \phi)$	Normal distribution with mean μ and precision ϕ
$\mathbf{G}(\cdot \mid a,\ b)$	Gamma distribution with shape a and rate b, which has a mean of a/b

Abbreviations

Here are the abbreviations of terms used in the study.

PSBP	Probit stick-breaking process
BNP	Bayesian nonparametric
OM	Operations management
C2k	Cargo 2000 standards

Acknowledgements

The authors are gratefully for the generous support of industry sponsors Dr. Prof. Rod Franklin and Michael Webber during the years. We greatly appreciate them for making the data available to the research, for sharing with us their rich knowledge of the industry and patiently waiting during the long research process.

1

Introduction

Global trade has grown considerably in recent decades; many companies now have overseas facilities and supply chain partners. International cargo logistics management thus plays an increasingly important role in the global economy. As one of the speediest transportation means, air transport delivers high quality products at competitive prices to customers worldwide. Indeed, air cargo transports goods worth in excess of $6.4 trillion annually. This is approximately 35% of world trade by value (IATA, 2014). This industry, including express traffic, is forecast to grow at an average 5.2% annual rate in the following two decades, from 202.4 billion RTKs (revenue tonne-kilometers) in 2011 to more than 558.3 billion RTKs in 2031 (Crabtree et al., 2012). However, attention paid to this industry is surprisingly little: air cargo industry '.. has remained the poor cousin to the more glamorous passenger side of the business (passenger air transport industry)' (Morrell, 2011).

The consequences of this neglect are significant as the service level of cargo transport has become firms' big concern. In cargo logistics, a key (service) performance measure is *transport risk* (or delivery reliability), defined as the deviation of the

actual arrival time from the planned arrival time,

$$\text{transport risk} = \text{actual arrival time} - \text{planned arrival time}.$$

Neither earliness nor tardiness is desirable for customer and freight forwarders. While tardiness causes delay in production and product/service delivery to all downstream customers, earliness incurs additional storage and handling costs. Extreme risks, such as more than 48 hour delays or more than 24 hours earliness, is defined as *(transport) disruption risks*, because they severely impact the operations of the customers and the freight forwarders. To distinguish disruption risks from the routine deviations within a day, we refer to the latter as *recurrent risks*. According to a 2011 PRTM survey, 69% of companies named improving delivery performance as their top supply chain management strategy. In a 2010 report of Infosys, "carrier delays and non-performance on delivery" is ranked as the top 1 risk in the logistics industry. Furthermore, in a 2014 survey conducted by International Air Transport Association (IATA) to major freight forwarders and their customers, low reliability is perceived as the second most important factor (next to transportation cost).

In this paper, we study the transport risks of international air cargo based on a half-year of air cargo data between 2012 and 2013, provided by a leading forwarder on 1336 routes served by 20 airlines. Using a Bayesian nonparametric (BNP) model – the Probit stick-breaking (PSBP) mixture model – we obtain accurate estimates of transport risk distributions and disruption risk probabilities. Our model provides a tool for the forwarder to offer customized price and service quotes. It can also generate baseline airline performance to enable fair supplier evaluation. Furthermore, the method allows us to separate recurrent risks from disruption risks. This is important, because hedging strategies for these two kinds of risks are often drastically different.

We make several contributions to the Operations Management (OM) literature as outlined below.

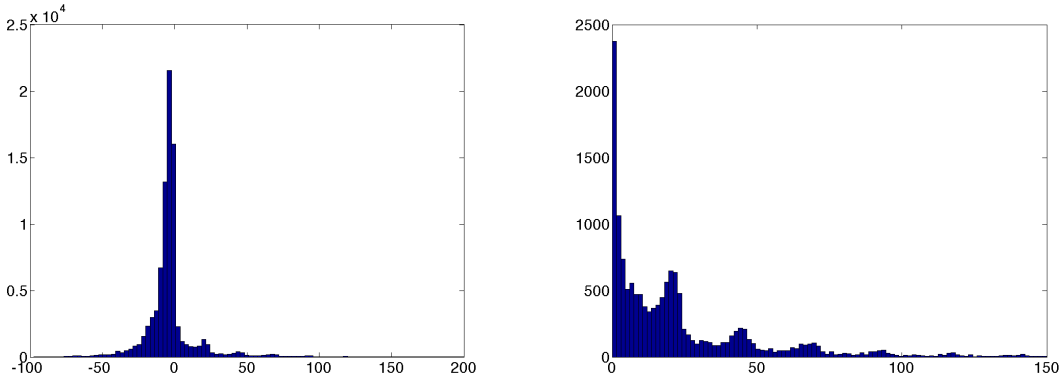

FIGURE 1.1: Histograms of transport risk (hours)

Empirical Air Cargo Transport Risk Distribution

First, our work appears to be the first empirical study of global air logistics in the supply chain literature. One interesting phenomenon observed from the data is that the distribution of transport risk, conditional on predictors, is a *multimodal distribution*, as shown in Figure 1.1. The left side of Figure 1.1 is the empirical distribution of transport risks of all shipments observed in the data (almost 90 thousand shipments), which, clearly, is a non-symmetric, long-tail distribution with several bumps at the distribution's positive part. To better observe the bumps, we only plot the data that falls in the range (0, 150) on the right side of Figure 1.1. Here, we can see clearly that big bumps concentrate around days (at 24 hours, 48 hours, and 72 hours, etc.) and small bumps between days. These systematic peaks are largely due to the fact that a cargo that failed to be loaded onto its scheduled flight was loaded onto a flight on the same route later. The scheduled gap between flights, which depends heavily on the route, for example, is usually around 24 hours for international flights and $4 \sim 6$ hours for domestic flights. The time gaps between scheduled flights thus transfer to the gaps between different peaks in the conditional distribution of transport risk to form a multimodal distribution; see Section 4 for more detail.

Previous empirical studies primarily focus on domestic passenger flight arrival or

departure delays; see Deshpande and Arikan 2012 for a review. (Note that delay or lateness is the positive part of transport risk, because earliness is usually not a concern for passenger flights.) Most of this literature assumes the delays follow unimodal distributions. Under this assumption, most works, such as Shumsky (1995) and Mueller and Chatterji (2002), adopt the classic ordinary least square linear regression (OLS for short) for delay estimation. However, our above multimodal observation indicates that OLS is unsuitable for the air cargo transport risk assessment and prediction. This is because OLS is built on the assumption that the distribution of a dependent variable, conditional on other predictors, is unimodal (most often Gaussian distribution). Hence, we need to develop new methodologies as described below.

The BNP Model

Our second contribution is methodological. To accommodate the multimodal feature in the empirical transport risk distribution, we apply a state-of-art Bayesian statistics tool – the BNP mixture model. To the best of our knowledge, no prior work has used related techniques in empirical OM, which so far predominantly applies frequentist statistics, such as OLS and maximum likelihood estimation (MLE) (see, e.g., Deshpande and Arikan (2012), Guajardo et al. (2012) and the references therein).

Bayesian statistics has experienced rapid development in the past two decades accelerated by ever-increasing machine computational power. Among these tools, BNP mixture models have become extremely popular in the last several years, with applications in fields as diverse as finance, econometrics, genetics, and medicine (refer to Rodriguez and Dunson (2011) for references). A nonparametric mixture model can be expressed as follows: in the case where we are interested in estimating a single distribution from an independent and identically distributed ($i.i.d$) sample

$y_1, \cdots, y_n$, observations arise from a convolution

$$y_j \sim \int k\left(\cdot \mid \boldsymbol{\psi}\right) G\left(\mathrm{d}\boldsymbol{\psi}\right)$$

where $k\left(\cdot \mid \boldsymbol{\psi}\right)$ is a given parametric kernel indexed by $\boldsymbol{\psi}$ (we use bold symbol to indicate vector), and G is a mixing distribution assigned a flexible prior

$$G\left(\boldsymbol{\psi}\right) = \sum_{l=1}^{L} \omega_l \delta_{\boldsymbol{\psi}_l}, \text{ where } \sum_{l=1}^{L} \omega_l = 1$$

and L could be finite or infinite. For example, assuming that G follows a Dirichlet process (DP) prior leads to the well known Dirichlet process mixture (DPM) model (Escobar and West, 1995).

For our application, we adopt a specific BNP model – the PSBP mixture model, which was formally developed in Rodriguez and Dunson (2011). This method is known for its flexibility, generality, and, more importantly, computational tractability. Meanwhile, PSBP leads to consistent estimation of any conditional density under weak regularity conditions as shown in Pati et al. (2013). Rodriguez et al. (2009) use this technique to create a nonparametric factor model to study genetic factors predictive of DNA damage and repair. Chung and Dunson (2009) apply this tool to develop a nonparametric variable selection framework to a data set from the Insulin Resistance Atherosclerosis Study (IRAS). Our model is designed to capture the transport risk distribution characteristics in all ranges, covering both recurrent and disruption risks.

To demonstrate the value of PSBP, we compare our transportation risk estimation with that obtained from the OLS model. We show that the two methods deliver dramatically different results. For instance, OLS fails to capture the critical roles airlines play in transport service levels. More importantly, the OLS predictions underestimate disruption risks, which can result in insufficient risk management strategies.

Data-Driven Risk Assessment Tool

Our method suggests a powerful and general tool to help supply chain risk assessment, a topic that has not received the attention it deserves. In particular, while supply chain risk management is gaining increasing attention from both practitioners and academics, a recent McKinsey & Co. Global Survey of Business Executives shows that "nearly one-quarter firms say their company doesn't have formal risk assessment, and almost half lack company-wide standards to help mitigate risk." Indeed, as articulated in Van Mieghem (2009), managing risk through operations contains 4 steps: (1) identification of hazards; (2) risk assessment; (3) tactical risk decisions; (4) implement strategic risk mitigation or hedging. These four steps must be executed and updated recurrently. Among the four steps, step 1 is more experience and context based, which typically involves information from anecdotal records or long experience with the specific business processes. Step 4 is more action-based, requiring detailed organizational design and information systems to carry out the hedging strategies developed in step 3. These two steps may not need quantitative methods. Steps 2 and 3, on the other hand, require rigorous analysis and quantification, and therefore call for analytical research. While most of the supply chain risk management literature focuses on the third step, which involves developing strategies for reducing the probabilities of negative events and/or their consequences should they occur, this paper focuses on step 2 – risk assessment.

Risk assessment can be decomposed into estimating two somewhat distinct components: (1) risk likelihood, i.e., "the probability that an adverse event or hazard will occur" and (2) risk impact, i.e., "the consequences of the adverse event" (Van Mieghem, 2009). The long-term expected risk is the integration of these two parts. Though scarce, we noticed a distinguished work by Kleindorfer et al. (2003) on assessing risk impact (part (2)) of catastrophic chemical accidents using data

collected by the Environmental Protection Agency. Kleindorfer and Saad (2005) present a conceptual framework for risk assessment and risk mitigation for supply chains facing disruptions. Different from these studies, our work focus on using statistical methods to accurately estimate the risk likelihood (part (1)), which calls for more advanced scientific computation and analysis tools.

Correctly identifying hazards and assessing risk has important implication for the effectiveness of alternative management policies(Cohen and Kunreuther, 2007), and our study shows that a careful risk assessment is critical to (1) developing tailored services for customers (i.e. shippers) of different types; (2) improving the operations efficiency of companies (i.e. forwarders), especially, in our case, when risks change dramatically depending on the different business situations and available choices.

The transport risk studied in this paper resembles the deviation between planned yield (capacity) and actual yield (capacity); the random yield/capacity risks problem in manufacturing is studied by many authors; see, e.g., Wang et al. 2010, Federgruen and Yang 2009. Also, the transportation disruption risk is an important type or component of random supply disruption risks considered by Tomlin 2006, etc. While all these authors focus on risk mitigation strategies assuming a particular risk distribution, such as a Bernoulli distribution for disruption risks, the Bayesian PSBP mixture model introduced here can be used to generate empirical random yield distributions and disruption probabilities, when data are available.

Finally, our study serves as a stepping stone to deeper studies in air cargo transport industry, or more generally, the transportation industry, which generates tons of data everyday yet lacks proper techniques for data analysis. According to a 2011 McKinsey report (Manyika et al., 2011), in the transportation and warehousing sector, the main focus of our paper, IT intensity is among the top 20% and data availability is among the top 40% of all sectors, but the data-driven mind-set is merely at the bottom 20%. The authors' communication with leaders in this industry, from

whom we get the data supporting this research project, confirms this situation, "... we have plenty of data, or we could say we have all the data possible, but we don't know how to use the data...".

The reminder of the paper is organized as follows: after a brief review of our research motivation and contribution, in Section 2, we give a general and brief introduction of the air cargo logistics industry, the existing problems faced and the data we used for this study; in Section 3, we describe exploratory analysis to lead to formal model selection, and we introduce the PSBP mixture model and the algorithm for posterior Gibbs sampling; in Section 4 we explain the results; in Section 5 we propose several applications of using our model to facilitate making more efficient operational strategies. In Section 6, we conclude the paper with existing problems and future directions.

2

Industry Background, Data Source, and Research Questions

Though a crucial part of global operations, the air cargo industry is less known to the public partly because it is always behind the scene. For this reason, in order to understand our model and analysis, it is necessity to provide brief background of the industry, which also explains the initial motivation for the industry to develop a standardized *Cargo 2000* process. We use *Cargo 2000* as our data source.

2.1 Service Chain Structure

First, let's take a look at the shipping process. Generally speaking, the completion of each single air cargo transport involves four parties: *shippers* (e.g., manufacturers), *freight forwarders* (*forwarder* in short), *carriers* (airlines) and *consignees* (e.g., downstream manufacturers or distributors); see Figure 2.1 for an illustration. These four parties form a chain structure, usually called the air transport supply chain. Specifically, a shipper initiates a shipping transaction by calling the local branch office of a forwarder company defining "(1) origin/destination; (2) collection/delivery

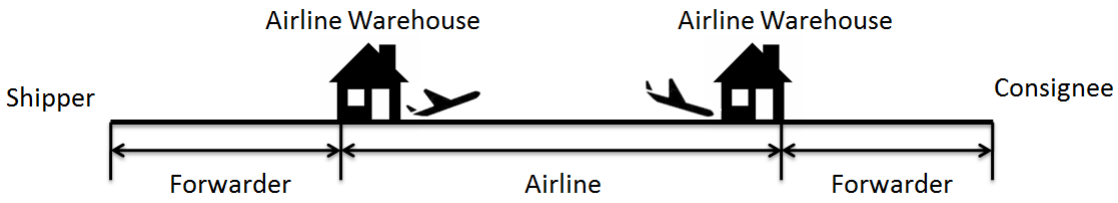

FIGURE 2.1: Cargo flow

date; (3) shipment details (cargo pieces, weight & volume); (4) shipper/consignee information; (5) product/shipping service required" (IATA, 2014). The next step is for the forwarder to create a route map (shipping plan) to meet the shipper's requirements that the forwarder and carrier are capable to conduct. After successfully creating a route map, the forwarder picks up cargoes from the shipper at the required time and consolidates cargoes sharing the same route together if applicable, then sends cargoes to a certain airline at an origin airport. The airline takes charge of cargoes until arriving at the destination airport. An airline might use a direct flight or $2 \sim 3$ connected flights based on the route map to ship cargoes to the destination. Just as passengers usually don't change airlines for connecting flights, each cargo shipping is also conducted by only one airline, specified in the route map. After picking up cargoes from the airline at destination airport, the same forwarder[1], which accepts cargoes at the origin, delivers them to consignees.

Cargoes shipped could be owned by the shipper, or the consignee, or a third party if applicable. Moreover, the payer of shipping fees can be any of these three. To simply terms, we refer to both the shipper and the consignee as the "customers". Except sending cargoes to a forwarder, there are several alternatives a customer could choose from, such as sending cargoes to an airline directly or to an integrator. In reality the majority of customers choose forwarders, constituting more than 90% of air cargo volume. A forwarder provides many value-added services besides air trans-

[1] A forwarder has branches all over the world so as to guarantee accepting/delivering cargoes everywhere

port, including cargo pick-up at origin, cargo storage, import-export documentation (e.g customer clearance) preparation, cargo delivery at destination etc. Moreover, since a forwarder can consolidate cargoes and thus achieve lower shipping rates from airlines, it is more economical for the customers. As such, a forwarder is a service provider for its customers.

On the other hand, a forwarder also uses airlines as service providers. Upon receiving a shipping request, a forwarder would send a booking request to several airlines, and choose the most economic one that matches its promises to the customer. More often, a forwarder allocates against its pre-booked shipping capacity. A large forwarder typically reserves a certain percentage (e.g. 30%) of the total space on routes from almost all possible airlines, including both passenger airlines and cargo airlines. Except the guaranteed capacity commitments made on part of airlines, the forwarder also get pre-fixed special shipping rates in contract, see Gupta (2008). This kind of capacity-rate contract is made between the forwarder with each airline every several months or even one year. The exact volumes and prices agreed on are determined by many factors, such as the popularity of the route, holiday season time, and the relationship between the two parties.

2.2 *Cargo 2000* (C2K) Standards

To compete against integrators'[2] reliable, time-definite transport services, *Cargo 2000* (C2K) was founded by a group of leading airlines and freight forwarder companies, "IATA Interest Group", in 1997 under the auspices of IATA. This initiative was designed to enable industry-wide participants to "provide reliable and timely delivery shipments through the entire air transport supply chain" (C2K MOP Executive Summary IATA 2014). Specifically, they developed a system of shipment planning

[2] Freight integrators are transport service providers who arrange full load, door-to-door transportation by selecting and combining without prejudice the most sustainable and efficient mode(s) of transportation, such as DHL, UPS etc.

and performance monitoring for air cargo based on common business process and milestones definition. Currently C2K is composed of more than 80 major airlines, forwarders, ground-handling agents, etc (see Figure A.1 for the current members of C2K), and aims to improve airfreight value through industry collaboration. C2K Quality Management System is implemented with two different scopes: Airport-to-Airport (A2A) and Door-to-Door (D2D). In this paper, we focus on the A2A level shipments due to data constraints. Because A2A is an essential element of D2D shipments, a good understanding of A2A shipments is an important starting point of research of D2D shipments.

Next, we explain how C2K is used to create a shipping plan, and more importantly, how airlines and forwarders achieve to "monitor, control, intervene and repair" (IATA, 2014) each shipment in real-time.

2.2.1 Plan

After a carrier has confirmed requested capacity on planned flights, it creates an A2A route map (RMP) and shares it with the forwarder through their common data management platform. A RMP describes the path the freight shipment follows, including flight information as well as milestones and the latest-by time for the fulfillment of the milestones along the transport chain. See Table A.2 and and Figure A.2 in Appendix for an illustration a RMP and milestones. If a customer agrees on the plan, the RMP is set alive. Otherwise, modifications will be made until agreement is achieved. Essentially, each route map is a combination of a station profile and milestones. Station profile, which contains information on the duration for completion of each process step, are kept by forwarders and carriers. The milestones are defined by the C2K Master Operating Plan (MOP).

2.2.2 Monitor, Control, Intervene and Repair

After a route map is issued, the actual shipping process is then automatically monitored against this route map. The completion of every milestone triggers updates on both the airline's as well as the forwarder's IT systems. Any deviation from the plan triggers an alarm, which allows for corrections to be taken by the responsible party in order to bring the shipment back on schedule. If necessary, a new RMP is made accordingly for the remaining transport steps. Meanwhile, an exception record is entered into the system recording the necessary information such as time, location, and reasons. See Table A.2 in Appendix for an illustration.

2.2.3 Report

At the end of the shipment process, a report, including whether or not the delivery promise was kept and which party was accountable for the failure, is generated. This allows the customers to directly compare the performance of their C2K enabled forwarders, carriers and logistics providers.

2.3 Forwarder's Frustration and Our Objectives

However, even with the help of highly integrated IT systems, which ensures real-time information sharing and industry-wide collaboration among supply chain parties (i.e., forwarders, airlines, customers) after exceptions, the service level is still not satisfying, as mentioned at the beginning of the paper. When a "poor" service happens, the forwarder, as the customer facing service provider, is the recipient of customer blames/complaints, and more importantly, faces the risk of losing customers. On the other hand, a forwarder has no actual control over the A2A part of the service, which is the most uncertain part during the entire shipping process. Hence, two challenging questions for the *forwarder* to solve are: (1) how to predict transport risks so as to prepare for risks and inform customers in advance and (2) how to improve transport

Table 2.1: Potential predictors

variable	description
cargo-related variables	
route	an origin-destination airport pair combination (captures all the fixed effects on a particular route).
month	month when the shipping is finished
cargo weight	total weight of the cargo (kilograms)
cargo volume	total volume of the cargo (cubic meters)
service-related variables	
airline	the airline transported the cargo
number of legs	number of connecting flights taken to arrival at destination
planned duration	total time (days) planned to take to finish the transport
initial deviation	deviation (days) between actual and planned check-in time at airline origin warehouse

reliability in each route by selecting the right supplier? We aim to help forwarders to address these questions in this paper.

Specifically, consider a customer comes to the forwarder for air cargo shipping with a fixed route (origin-destination pair) in mind, time of shipping, weight and volume of cargo. We aim to enable the forwarder to give a distribution of transport risk conditional on all the predetermined cargo-related variables (route, month, cargo weight/volume) and selectable service variables (airline, number of flight legs, planned duration, initial deviation time) with 95% uncertainty interval. See Table 2.1 for more detailed descriptions of these variables. Based on this information, the customer will be able to find a favorable combination of selectable service variables depending on their own cost/utility function. Meanwhile, the forwarder will be able to provide different price quotes for different services targeting different customers, which can help yield larger profit.

Next, we elaborate how the above mentioned cargo- and service-related variables affect the transport risk.

2.3.1 Effect of Cargo-Related Variables

1. Route: the service level differs dramatically from route to route depending on (1) the demand and available supply of air transport service on that route and (2) the congestion level and infrastructure quality at origin and destination airports, such as whether the origin/destination is a hub or in an emerging market. Since we are not testing hypotheses regarding the relationship between these factors (hub, region, demand etc) with transport service levels, we do not separate these factors. Instead, we use a route-level effect to absorb the effect of all these factors.

2. Month: demand (holiday shipping etc) and weather (winter snow etc) both have a seasonal trend, which results in different perceived air cargo transport service levels in different months. We used the month when the transport is completed as the predictor; since shipments only take 1.7 days to finish on average, essentially identical results would be achieved using the month of transport start.

3. Cargo weight and volume: each flight has a capacity constraint on the maximum weight and volume. On one hand, compared to small cargoes (measured in weight or volume), larger cargoes are more likely to fail to be loaded onto the scheduled flight due to (1) airlines' overselling capacities and (2) any slight changes of currently available capacity, such as more check-in luggage from passengers. Thus, we expect to observe worse services for larger cargoes ceteris paribus. On the other hand, larger cargoes are usually more valuable than smaller cargoes and thus may have higher transport priority and thus a more reliable service. Clearly, it is not easy to disentangle these two effects, but our analysis can help reveal which one is more dominant.

2.3.2 Effect of Service-Related Variables

1. Airline: because different airlines use different sizes of flights, booking strategies (e.g. portion of over-booked capacity to the total capacity), scheduling strategies (e.g. percentage of cushion added into the total shipping schedule) etc., airlines affect the distribution of transport risk in a complex way. In addition, airlines may provide varying service levels across routes depending on factors such as whether this airline has a hub along the route, the nationality of the airline, etc, and hence we added the interaction of airline and route into the model.

2. Number of legs: number of legs increases the probability for a cargo to miss connecting flights, so it is a strong predictor of transport risk. Although in many routes, after choosing a particular airline, the number of legs is simultaneously determined, we do observe a large amount of routes on which one airline offers services with different legs (usually both direct and two-leg services on the same route). So we choose to add the number of legs as one predictor and also a changeable factor the customer can choose.

3. Planned duration: conditional on route, airline and number of legs, we still observe planned duration differs greatly from one another. This reflects the fact that cushions are added into the route map since the air flying time should be nearly constant. In principle, the larger the cushion the lower the delay probability. For example, given the first flight is delayed, if the cushion (connecting time) is long enough, the cargo can still catch the next flight, however, if the cushion is small, the cargo might miss the second flight resulting in severely delayed final delivery. However, a larger cushion might reflect airline's private information of congested traffic and thus is a signal of possible delays. So

whether a longer duration (a larger cushion) would imply improved transport reliability is to be analyzed.

4. Initial deviation: if the cargo is sent to the airline earlier than scheduled, it could be loaded onto an earlier flight and vice versa.

2.3.3 Discussion: Other Potential Predictors

We note that there are other factors that may affect the risk distribution, such as price and weather. However, these are unobservable from the data set we have. This is why we only use the predictors explained above. Nonetheless, our model indirectly captures some important effect of these unobserved factors. For example, shipping price, which determines the service priority, is calculated based on cargo weight/volume, route, airline, number of legs and planned duration (speedy service or standard service). But the latter factors are all included in our model. So even though we don't observe price, our model captures the effect of shipping priority and class. Similarly, weather information, which heavily depends on geographic location and season, is partly included in the predetermined route and month variables. It is quite challenging to find more detailed weather information at each moment and each place for our international shipments in the half year time frame. However, if such data are available in the future, it will be straightforward to be added into our model.

2.4 Data and Summary Statistics

As mentioned before, our data are provided by one of the world leading freight forwarder companies. The data contain the company's C2K standard airfreight shipments from 2012 October to 2013 April (about half a year). Specifically, it contains historical records of real-time milestone updates, which are similar to the data shown in Table A.2. The other equally important parts are the route maps associated with

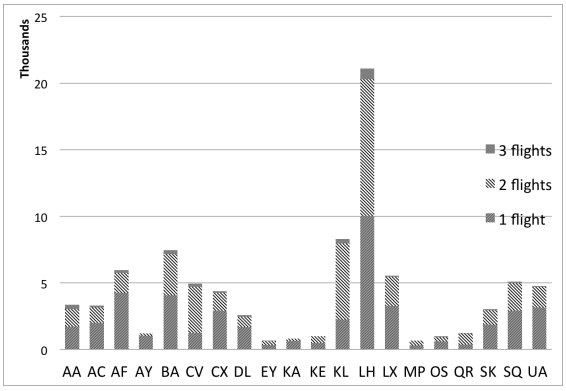

FIGURE 2.2: Number of shipments by each airline

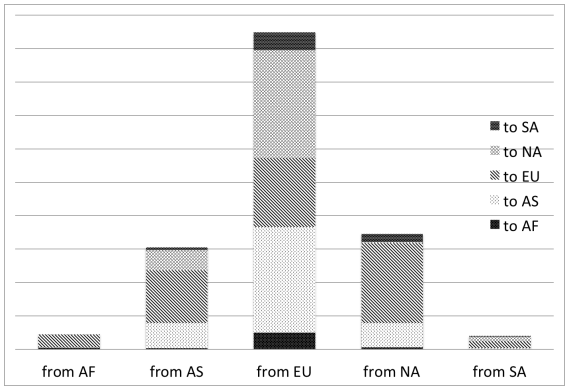

FIGURE 2.3: Number of shipments between continents

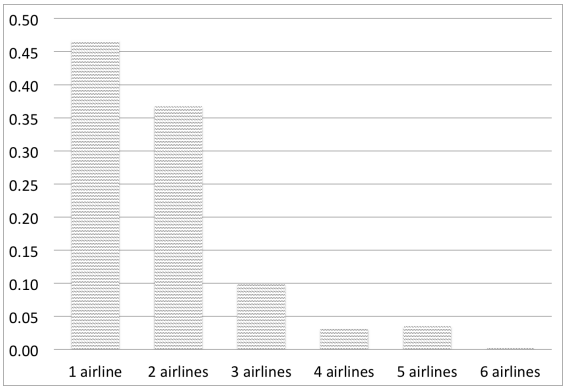

FIGURE 2.4: # of airlines faced by shippers on each route

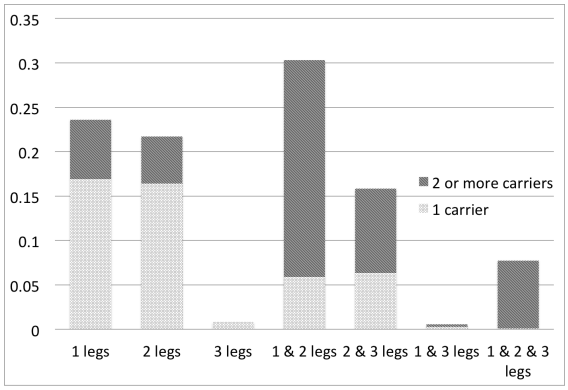

FIGURE 2.5: # of legs faced by shipper on each route

each shipment. Following the company's advice and also adopting industry standards, we use the last route map made before the shipment occurs as the baseline route map against which to measure and benchmark "performance vs. promise". After cleaning (see Appendix for detailed cleaning steps), the data we use for analysis include 86,149 shipments on 1336 routes operated by 20 airlines. The freights are shipped from 58 countries to 95 countries, see Figure 2.3 for the percentage of shipments between the five continents[3]. In Figure 2.2 is the number of shipments shipped by each airline, and the percentage of shipments by different number of legs.

Combining Figures 2.2 and 2.3 we can see why European airlines, such as Lufthansa

[3] AF: Africa; AS: Asia; EU: Europe; NA: North America; SA: South America

Table 2.2: Summary statistics

Dependent Variable						
	mean	std				
transport risk (hour)	-2.6	20.6				
Predictors						
Category Predictor						
	airline	route	airline-route	month	airline-leg2	airline-leg3
dim	20	1336	588	7	20	16
Continuous Predictor						
	dev_{start} (day)	dur (day)	$\log(wgt)$ (kg)	$\log(pcs)$ (cbm)		
mean	-0.327	1.75	4.91	1.29		
std	0.648	1.30	2.4	1.43		

and KLM, play a significant role in the data. Figure 2.4 shows the number of airlines each shipment is choosing from, from which we can see that more than 50% of shipments are transported on routes served by more than 1 airline. Figure 2.5 depicts the choices between legs each shipment is facing. There are more than 50% of shipments transported on routes where services of different legs are available. For example, around 30% shipments (the fourth column) are on routes served both by direct flight and 2-leg service. Figures 2.4 and 2.5 indicate that a majority of shipments are facing the choice between number of legs or airlines or both, in which situation a careful inspection and assessment of different choices can help achieve a superior utility if service level vary significantly across choices.

Table 2.2 provides the summary statistics of the dependent variable, transport risk, and potential predictors explained in Table 2.1.

2.4.1 Exception Records

The creation of the exception codes is meant to facilitate (1) finding root causes of delays and (2) identifying parties accountable for failures. Unfortunately, however, the exception information recorded from the data is not helpful in regard to these

two goals. (This fact was also confirmed by the company.) First, the data missing rate is high. Only less than 8% percent of the milestones delayed and less than 10% of milestones delayed for more than 1 day have exception information recorded. Second, the exception codes used are highly ambiguous. For example, the most frequently appearing code is "COCNR", which means the carrier hasn't received the cargo. However, why the cargo is not received and where the cargo could be are not included in the message. As a result, we do not use exception data for our analysis.

3

Model

We have discussed the multimodal distribution of transport risk at the beginning of the paper by showing the empirical distribution of the whole data set (see Figure 1.1). This multimodal feature is not only present at the whole data level but also at the granular level, such as each route or route-airline level. See Figure 3.1 for the empirical distribution on two sample routes served by two airlines. In order to make accurate predictions and inferences based on such data, the first step is choosing a model flexible enough to fit the data well. Usual choices of models for multimodal

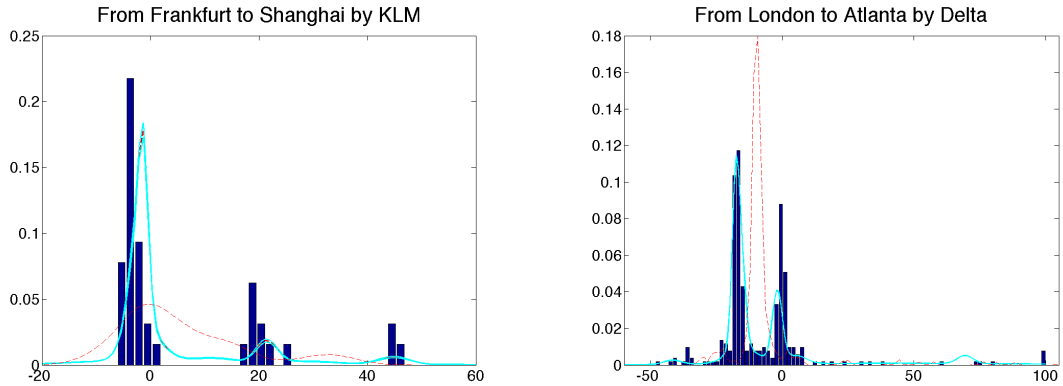

FIGURE 3.1: Sample routes

data rely on mixtures, e.g., mixtures of Gaussian kernels, which are known to provide an accurate approximation to any unknown density.

We cannot rely on simple mixture models, as we are investigating the distribution of transport risks conditional on the cargo-related and service-related variables, including both categorical and continuous predictors. This leads to a problem of ***conditional distribution estimation***. One stream of literature on flexible conditional distribution estimation uses frequentist methods. Fan et al. (1996) proposed a double-kernel local linear approach, and related frequentist methods have been considered by Hall et al. (1999) and Hyndman and Yao (2002) among others. The other popular choice is a BNP mixture model. The seminal work of Muller et al. (1996) proposed a Bayesian approach to nonlinear regression, in which the authors modeled the joint distribution of dependent variable and predictors using a DPM of Gaussians (Lo 1984; Escobar and West 1995). This type of approach relies on inducing a model for the conditional distribution of the response through a joint model for the response and predictors. Although such joint models are provably flexible, in practice they can have clear disadvantages relative to models that directly target the conditional response distribution without needing to model the high-dimensional nuisance parameter corresponding to the joint density of the predictors. Such disadvantages include treating the predictors as random, while they are often designed variables (e.g., it seems unnatural to consider route or airline as random), and relatively poor practical performance in estimating the conditional and prediction.

In this article, we instead focus on direct modeling of the unknown conditional distribution of the delay y given predictors $\mathbf{x} = (x_1, \cdots, x_p)' \in \mathcal{X}$ ($\mathcal{X}$ is the sample space for the predictors $\mathbf{x}$) without specifying a model for the marginal of the predictors. In particular, we assume the delay data y arise from a convolution

$$y \mid \mathbf{x} \sim \int k\left(y \mid \boldsymbol{\psi}\right) G_{\mathbf{x}}\left(\mathrm{d}\boldsymbol{\psi}\right)$$

where $k\left(\cdot \mid \boldsymbol{\psi}\right)$ is a given parametric kernel indexed by parameters$\boldsymbol{\psi}$ (e.g., Gaussian), and the mixing distribution $G_{\mathbf{x}}$ is allowed to vary flexibly with predictors$\mathbf{x} \in \mathcal{X}$. The general form that is typically taken in the BNP literature (refer to Rodriguez and Dunson (2011) for related references) lets

$$G_{\mathbf{x}} = \sum_{l=1}^{L} \omega_l\left(\mathbf{x}\right) \delta_{\boldsymbol{\psi}_l(\mathbf{x})}, \text{ where } \sum_{l=1}^{L} \omega_l\left(\mathbf{x}\right) = 1$$

the atoms $\{\boldsymbol{\psi}_l\left(\mathbf{x}\right) : \mathbf{x} \in \mathcal{X}\}_{l=1}^{L}$ are $i.i.d$ sample paths from a stochastic process over $\mathcal{X}$, and $\{\omega_l\left(\mathbf{x}\right), \mathbf{x} \in \mathcal{X}\}$ are predictor-dependent probability weights that sum to one for all $x \in \mathcal{X}$. The above form is too general to be useful and it is necessary to make some simplifications for practical implementation. One common possibility is to introduce predictor dependence only in the $G_{\mathbf{x}}$ atoms, $\phi_l\left(x\right)$, while keeping weights, $\omega_l\left(\mathbf{x}\right) = \omega_l$, independent of predictors $\mathbf{x}$. However, this approach tends to have relatively poor performance in our experience, including with the flight delay data, compared with models that instead fix the atoms, while allowing the weights to vary.

In our case, the peak locations of the dependent variable, transport risks, are almost constant (i.e. daily peaks for international shipments, and some additional few-hourly peaks for domestic shipments besides the daily peaks). However, the heights of the peaks change greatly along with $\mathbf{x}$ (e.g. route, airline, cargo-related variables etc). The height of each peak represents (roughly) the probability for the observation to fall into the kernel centered around that peak. For example, if conditional on certain $\mathbf{x}_1$, the peak around 24 hours is relatively high, then a shipment, conditional on $\mathbf{x}_1$, has a considerable large probability of being delayed

for one day. While if conditional on certain $\mathbf{x}_2$, there is only one peak around 0 high and visible, then a shipment, conditional on $\mathbf{x}_2$, probably arrives close to the planned arrival time. So, in our context, to find out how the height of each peak changes with $\mathbf{x}$ is of central interest.

Inducing dependence structure in the weights can be difficult and lead to complex and inefficient computational algorithms, limiting the applicability of the models. The PSBP mixture model we use in this paper has distinct advantages over previous formulations in terms of computational tractability and consistency under weak regularity conditions. In this Section, we will explain the model, posterior computation algorithm and prior elicitation criteria, and we conclude this Section with model checking and selection.

3.1 Bayesian Probit Stick-breaking Process

As we have explained before, for mixing prior

$$G_{\mathbf{x}} = \sum_{l=1}^{L} \omega_l(\mathbf{x})\, \delta_{\boldsymbol{\psi}_l(\mathbf{x})}, \text{ where } \sum_{l=1}^{L} \omega_l(\mathbf{x}) = 1$$

and L is finite or infinite. We use constant atoms, $\boldsymbol{\psi}_l(\mathbf{x}) = \boldsymbol{\psi}_l \ \forall \mathbf{x} \in \mathcal{X}$, which are *i.i.d.* distributed from centering measure G_0. Stick-breaking weights are defined as $\omega_l = u_l \prod_{p<l}(1-u_p)$, where the stick-breaking ratios are independently distributed $u_l \sim H_l$ for $l < L$ and $u_L = 1$. In the baseline case in which there are no predictors, ***Probit*** stick-breaking weights are constructed as[1]

$$u_l = \Phi(\gamma_l), \ \gamma_l \sim \mathsf{N}(\mu, \phi)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function for the standard normal distribution. This in turn implies that the probability weight on the lth kernel can

[1] For $x \sim \mathsf{N}(x \mid \mu, \phi)$, the probability density function is $f(x) = \sqrt{\frac{\phi}{2\pi}} \exp\left\{-\frac{\phi}{2}(x-\mu)^2\right\}$

be expressed as

$$\omega_l = \Phi(\gamma_l) \prod_{p<l} (1 - \Phi(\gamma_p))$$

For a finite L, the construction of the weights ensures that $\sum_{l=1}^{L} \omega_l = 1$. When $L = \infty$, $\sum_{l=1}^{\infty} \omega_l = 1$ almost surely (see Rodriguez and Dunson 2011). The use of Probit transformation to define the weights allows researchers to restate the model using normally distributed latent variables, facilitating computation via data augmentation Gibbs sampling algorithms while also making model extensions to include additional structure (e.g,. predictors) straightforward. Additionally, the Probit transformation induces a natural scale in the transformed weights that simplifies prior elicitation.

In order to make $\omega_l(\mathbf{x})$ dependent on predictors $\mathbf{x}$, we replace γ_l with a function of $\mathbf{x}$, $\{\gamma_l(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}\}$, thus incorporating predictors $\mathbf{x}$ into the construction of $\{\omega_l(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}\}$. Particularly, we add a linear regression structure into the latent random variables $\gamma_l(\mathbf{x})$, where $\mathbf{x}$=\{airline (a), route (r), month (m), number of legs (leg), initial deviation (dev_{start}), planned duration (dur), cargo weight (wgt), cargo number of pieces (pcs)\} (as specified in Table 2.1)[2]:

$$\begin{aligned}
\omega_l(\mathbf{x}) &= \Phi(\gamma_l(\mathbf{x})) \prod_{p<l} (1 - \Phi(\gamma_p(\mathbf{x}))) \\
\gamma_l(\mathbf{x}) &= \theta_l^1 + \theta_a^2 + \theta_r^3 + \theta_{(a,r)}^4 + \theta_m^5 + \theta_{leg}^6 + \theta_{(a,leg)}^7 + f_1\left(dev_{start} \mid \boldsymbol{\theta}^8\right) \\
&\quad + f_2\left(dur \mid \boldsymbol{\theta}^9\right) + f_3\left(\log(wgt) \mid \boldsymbol{\theta}^{10}\right) + f_4\left(\log(pcs) \mid \boldsymbol{\theta}^{11}\right) \qquad (3.1)
\end{aligned}$$

where $\{\theta_l^1\}$ controls the baseline probability of latent class l, $\{\theta_a^2\}$ controls the baseline heterogeneity of airline a, $\{\theta_r^3\}$ controls the heterogeneity of route r, $\left\{\theta_{(a,r)}^4\right\}$ represents dependence of the weights on possible interactions between airlines and routes, and the meanings of $\{\theta_m^5\}$, $\left\{\theta_{leg}^6\right\}$,$\left\{\theta_{(a,leg)}^7\right\}$ are similar. Besides, f_1, f_2, f_3 and f_4 are spline functions expressed as a linear combination of B-splines of degree 4,

[2] In this paper we use superscript as an index rather than the exponent of the parameter

where the knots of dev_{start} are [-3, -2, -1, 0, 1, 2, 3], the knots of dur are [1, 2, 4, 6, 8, 10], the knots of $log(weight)$ are [2, 4, 6, 8] and the knots of $log(pcs)$ are [1, 3, 5]. Here we use the logarithm form of cargo weight (wgt) and number of pieces (pcs) as the predictors, since the original distributions are highly skewed. To ensure identification of the parameters, we let $\theta_1^2 = \theta_1^3 = \theta_{(1,r)}^4 = \theta_{(a,1)}^4 = \theta_1^5 = \theta_{(1,leg)}^6 = \theta_{(a,1)}^7 = 0$ for all a, r and interactions in sample space $\mathcal{X}$. To retain conjugacy, we choose Gaussian priors for parameters $\Theta = \{\{\theta_l^1\}, \{\theta_a^2\}, \{\theta_r^3\}, \left\{\theta_{(a,r)}^4\right\}, \{\theta_m^5\}, \left\{\theta_{leg}^6\right\}, \left\{\theta_{(a,leg)}^7\right\}, \boldsymbol{\theta}^8, \boldsymbol{\theta}^9, \boldsymbol{\theta}^{10}, \boldsymbol{\theta}^{11}\}$

$$\theta_j^i \sim \mathsf{N}\left(\theta_j^i \mid \nu^i, \epsilon^i\right), \text{ for } i = 8, \cdots, 11 \text{ and } j = 1, \cdots, n(i)$$

where $n(i)$ is the number of B-spline basis used of predictor i. For the coefficients of 7 categorical predictors $\boldsymbol{\theta}^1, \cdots, \boldsymbol{\theta}^7$ (i.e. $\boldsymbol{\theta}^1 = \{\theta_l^1\}$ etc), we build a hierarchy, which enables information borrowing among parameters in one category

$$\begin{aligned}
\theta_l^1 &\sim \mathsf{N}\left(\theta_l^1 \mid \Phi^{-1}\left(\frac{1}{L-l+1}\right), \epsilon^1\right), \\
\theta_a^2 &\sim \mathsf{N}\left(\theta_a^2 \mid 0, \epsilon^2\right), \\
&\vdots \\
\theta_{(a,leg)}^7 &\sim \mathsf{N}\left(\theta_{(a,leg)}^7 \mid 0, \epsilon^7\right).
\end{aligned}$$

where $\epsilon^i \sim \mathsf{G}(c_i, d_i)$ for $i = 1, 2, \cdots, 7$. Here we use the specially designed prior of θ_l^1 to enforce the same prior baseline probability of each cluster $l = 1, 2, \cdots, L$. The specification of $\{(c_i, d_i), \text{ for } i = 1, 2, \cdots, 7\}$ and $\{(\nu^i, \epsilon^i), \text{ for } i = 8, 9, \cdots, 12\}$ is discussed in Subsection 3.4.

3.1.1 Gaussian Kernel

A mixture of a moderate number of Gaussians is known to produce an accurate approximation of any smooth density. Also motivated by computational tractability of

the Gaussian distribution (e.g, through conjugacy in posterior calculations), we specify the parametric kernel, $k(\cdot \mid \boldsymbol{\psi})$, of PSBP mixture model as Gaussian, $\mathsf{N}(\cdot \mid \mu, \phi)$, where $\boldsymbol{\psi} = (\mu, \phi)$. Recall that our mixture model takes the form:

$$y \mid \mathbf{x} \sim \int k(y \mid \boldsymbol{\psi}) G_{\mathbf{x}}(\mathrm{d}\boldsymbol{\psi})$$

We replace the kernel in the above equation with Gaussian and use the PSBP specified prior $G_{\mathbf{x}}$. Then the conditional distribution of y can be expressed in the simple form

$$y \mid \mathbf{x} = \sum_{l=1}^{L} \omega_l(\mathbf{x}) \mathsf{N}(y \mid \mu_l, \phi_l)$$

where atoms $\{\psi_l = (\mu_l, \phi_l), \forall l = 1, 2, \cdots, L\}$ are the normal mean and precision for the lth component density, and are $i.i.d$ samples from centering measure $G_0 = \mathsf{NG}(\zeta_\mu, \xi_\mu, a_\phi, b_\phi)$, a conjugate Normal-Gamma prior[3]

$$\begin{aligned} \mu_l &\sim \mathsf{N}(\mu_l \mid \zeta_\mu, \xi_\mu), \\ \phi_l &\sim \mathsf{G}(\phi_l \mid a_\phi, b_\phi). \end{aligned}$$

where $l = 1, 2, \cdots, L$. The specification of prior $\zeta_\mu, \xi_\mu, a_\phi$ and b_ϕ is discussed in Subsection 3.4.

3.2 Posterior Computation

3.2.1 Gibbs Sampling for Constant Atoms

First we focus on case when $L < \infty$. For each observation $y_j \mid \mathbf{x}$, (corresponding to replicate j conditional on $\mathbf{x}$, $j = 1, \cdots, n(\mathbf{x})$ if there are replicates, otherwise j is dropped if there is no replicates conditional $\mathbf{x}$), we introduce a latent indicator variable $s_j(\mathbf{x})$[4] such that $s_j(\mathbf{x}) = l$ if and only if observation $y_j \mid \mathbf{x}$ is sampled from

[3] For $x \sim \mathsf{G}(x \mid a, b)$, the probability density function is $f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$

[4] we use $s(\mathbf{x})$ and $s \mid \mathbf{x}$ interchangeably, and the same applies to other variables such as $y \mid \mathbf{x}$ and $y(\mathbf{x})$

mixture component l, $l = 1, 2, \cdots, L$. The use of these latent variables is standard in mixture models; conditional on the indicators, the *full conditional distribution* of the component-specific parameters, μ_l and ϕ_l here, is given by

$$p\left(\mu_l, \phi_l \mid \cdots\right) \propto G_0\left(\mu_l, \phi_l \mid \zeta_\mu, \xi_\mu, a_\phi, b_\phi\right) \prod_{(\mathbf{x},j)\ s.t.\ s_j(\mathbf{x})=l} \mathsf{N}\left(y_j \mid \mu_l, \phi_l\right)$$

where G_0 is the Normal-Gamma conjugate prior of μ_l and ϕ_l. Simplified by the conjugacy structure we use, the Gibbs sampling is carried out by

$$\mu_l \mid \cdots \sim \mathsf{N}\left(\mu_l \mid \left[\zeta_\mu + n_l\phi_l\right]^{-1}\left[\zeta_\mu\xi_\mu + h_l\phi_l\right], \xi_\mu + n_l\phi_l\right)$$

where $n_l = \sum_{\mathbf{x}\in\mathcal{X}} \sum_{j=1}^{n(\mathbf{x})} \mathbf{1}_{(s_j(\mathbf{x})=l)}$ and $h_l = \sum_{\mathbf{x}\in\mathcal{X}} \sum_{j=1}^{n(\mathbf{x})} y_j(\mathbf{x})\mathbf{1}_{(s_j(\mathbf{x})=l)}$. Similarly, the Gibbs sampling for kernel precisions ϕ_l is

$$\phi_l \mid \cdots \sim \mathsf{G}\left(\phi_l \mid a_\phi + \frac{n_l}{2}; b_\phi + \frac{1}{2}\sum_{\mathbf{x}\in\mathcal{X}}\sum_{j=1}^{n(\mathbf{x})}\left(y_j(\mathbf{x}) - \mu_l\right)^2 \mathbf{1}_{(s_j(\mathbf{x})=l)}\right)$$

3.2.2 Gibbs Sampling for Latent Indicators

Conditional on the component specific parameters and the realized values of the weights $\left\{\omega_l\left(\mathbf{x}\right), \forall \mathbf{x} \in \mathcal{X}\right\}_{l=1}^{L}$, the full conditional distribution for the indicators is multinomial with probability given by

$$\Pr(s_j(\mathbf{x}) = l|\cdots) \propto \omega_l(\mathbf{x})\mathsf{N}\left(y_j(\mathbf{x})|\mu_l, \phi_l\right),$$

which yields a simple form for the conditional probabilities that we use in sampling from the multinomial conditional distribution:

$$Pr(s_j(\mathbf{x}) = l|\cdots) = \frac{\omega_l(\mathbf{x})\mathsf{N}(y_j(\mathbf{x})|\mu_l, \phi_l)}{\sum_{p=1}^{L}\omega_p(\mathbf{x})\mathsf{N}(y_j(\mathbf{x})|\mu_l, \phi_l)}$$

3.2.3 Gibbs Sampling for Latent Auxiliary Variable

In order to sample the latent processes $\left\{\gamma_l\left(\mathbf{x}\right), \forall \mathbf{x} \in \mathcal{X}\right\}_{l=1}^{L}$ and the corresponding weights $\left\{\omega_l\left(\mathbf{x}\right), \forall \mathbf{x} \in \mathcal{X}\right\}_{l=1}^{L}$, we introduce a collection of conditionally independent

latent variables $z_{jl}(\mathbf{x}) \sim \mathsf{N}(\gamma_l(\mathbf{x}), 1)$ and define $s_j(\mathbf{x}) = l$ if $z_{jp}(\mathbf{x}) < 0$ for all $p < l$ and $z_{jl}(\mathbf{x}) > 0$. We have[5]

$$\begin{aligned}\Pr(s_j(\mathbf{x}) = l) &= \Pr\left(z_{jl}(\mathbf{x}) > 0, z_{jp}(\mathbf{x}) < 0 \text{ for } p < l\right) \\ &= \Phi\left(\gamma_l(\mathbf{x})\right) \prod_{p<l} \left\{1 - \Phi\left(\gamma_p(\mathbf{x})\right)\right\}\end{aligned}$$

independently for each j. This data augmentation scheme simplifies computation as it allows us to implement the following Gibbs sampling scheme

$$z_{jl}(\mathbf{x}) \mid \cdots \sim \mathsf{N}\left(z_{jl}(\mathbf{x}) \mid \gamma_l(\mathbf{x}), 1\right) \mathbf{1}_{\mathbf{\Omega}_l}, \ \forall l \leqslant \min\{s_j(\mathbf{x}), L-1\},$$

with

$$\mathbf{\Omega}_l = \begin{cases} \{z_{jl}(\mathbf{x}) | z_{jl}(\mathbf{x}) < 0\}, & \text{if } l < s_j(\mathbf{x}), \\ \{z_{jl}(\mathbf{x}) | z_{jl}(\mathbf{x}) \geqslant 0\}, & \text{if } l = s_j(\mathbf{x}) < L \end{cases}$$

where $\mathsf{N}\left(\mu, \phi\right) \mathbf{1}_\Omega$ denotes the normal distribution with mean μ and precision ϕ truncated to the set Ω.

3.2.4 Gibbs Sampling for Latent Processes

Conditional on the augmented variables $\{\mathbf{z}_j(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}\}$ ($\mathbf{z}_j(\mathbf{x}) = \{z_{j1}(\mathbf{x}), \cdots, z_{js(\mathbf{x})}(\mathbf{x})\}$), the full conditional posterior distribution of parameters $\Theta = \{\{\theta_l^1\}, \{\theta_a^2\}, \{\theta_r^3\}, \left\{\theta_{(a,r)}^4\right\}, \{\theta_m^5\}, \left\{\theta_{leg}^6\right\}, \left\{\theta_{(a,leg)}^7\right\}, \boldsymbol{\theta}^8, \boldsymbol{\theta}^9, \boldsymbol{\theta}^{10}, \boldsymbol{\theta}^{11}, \boldsymbol{\theta}^{12}\}$ and $\Upsilon = \{\epsilon^i, \forall i = =1, 2, \cdots, 7\}$, on which the latent processes $\{\gamma_l(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}\}$ are built, is given by

$$p(\Theta, \Upsilon \mid \cdots) \propto \left[\prod_{\mathbf{x},j} p\left(\mathbf{z}_j(\mathbf{x}) \mid \boldsymbol{\gamma}_j(\mathbf{x})\right)\right] p(\Theta) p\left(\Upsilon\right)$$

where $p\left(\Theta\right)$ is the prior distribution of Θ and $p\left(\Upsilon\right)$ is the prior distribution of Υ. The posterior sampling can be easily implemented by taking advantage of the normal

[5] with $s_i(\mathbf{x}) = L$ if $z_{ip}(\mathbf{x}) < 0$ for all $p \leqslant L-1$ for the finite L case

priors we use. Due to similarities of the Gibbs sampling schemes for coefficients in Θ as well as hyper-parameters Υ, here we only give updating schemes for two examples: one for coefficients $\{\theta_l^1\}_{l=1}^L \in \Theta$ and the other one for hyper-parameter $\epsilon^1 \in \Upsilon$. Interested readers can refer to the Appendix for exact equations used for Gibbs sampling of other parameters.

1. For θ_l^1 $(l = 1, 2, \cdots, L)$, the posterior Gibbs sampling follows normal distribution given by

$$\theta_l^1 \mid \cdots \propto \mathsf{N}\left(\theta_l^1 \mid \mu_{\theta_l^1}, \phi_{\theta_l^1}\right)$$

where $\mu_{\theta_l^1} = \left\{\Phi^{-1}\left(\frac{1}{L-l+1}\right) + \sum_{\mathbf{x}\in\mathcal{X}}\sum_j \left[z_{jl}\left(\mathbf{x}\right) - \Delta_{jl}\left(\mathbf{x}\right)\right]\mathbf{1}\left(s_j\left(\mathbf{x}\right) \geqslant l\right)\right\} / \left(n_l + 1\right)$,

$\phi_{\theta_l^1} = \left(n_l + \epsilon^1\right) / \left(n_l + 1\right)$, $n_l = \sum_{\mathbf{x}\in\mathcal{X}}\sum_j \mathbf{1}\left(s_j\left(\mathbf{x}\right) \geqslant l\right)$ and

$\Delta_{jl}\left(\mathbf{x}\right) = \left(\gamma_j\left(\mathbf{x}\right) - \theta_l^1\right)\mathbf{1}\left(s_j\left(\mathbf{x}\right) \geqslant l\right)$.

2. For ϵ^1 the posterior Gibbs sampling follows Gamma distribution given by

$$\epsilon^1 \mid \cdots \propto \mathsf{G}\left(\epsilon^1 \mid c_1 + \frac{L}{2}, d_1 + \frac{\sum_{l=1}^L \theta_l^1 \cdot \theta_l^1}{2}\right)$$

Data augmentation strategies of this kind allow for implementations that rely only on Gibbs samplers, rather than general MCMC schemes requiring simultaneous proposals of large numbers of parameters or rejection samplers that could generate even worse mixing issues by forcing us to sample one parameter at a time. In the case $L = \infty$, we can easily extend this algorithm to generate a slice sampler, as discussed in Papaspiliopoulos (2008). Alternatively, the results in Rodriguez and Dunson (2011) suggest that a finite PSBP with a large number of components ($30 \sim 40$, depending on the value of μ) can be used instead (Ishwaran and Zarepour 2002). So we use $L = 50$ as the number of components in this paper, the provides a conservative upper bound as many of these components may not be utilized.

3.3 Label Switching Moves

In general, in the conditional method, the Markov chain Monte Carlo algorithm has to explore multimodal posterior distributions. Therefore, we need to add label-switching moves, which assist the algorithm in jumping across modes. This is particularly important for large data sets, where the modes are separated by areas of negligible probability. We use the framework developed in Papaspiliopoulos and Roberts 2008 to design our label switching moves. These label switching moves greatly improved the convergence of the chain. See Appendix for explanation and algorithms of the moves.

3.4 Prior Elicitation

Consider first eliciting hyper-parameters $\{\zeta_{\mu_l}\}_{l=1}^{L}$ and $\{\xi_{\mu_l}\}_{l=1}^{L}$ corresponding to the location of the Gaussian components and a_ϕ and b_ϕ corresponding to their precisions. These hyper-parameters need to be chosen to ensure that the mixture spans the expected range of observed values with high probability. In practice, we have all prior means $\{\zeta_{\mu_l}\}_{l=1}^{L}$ equal to the global mean (or global median) of all observations in the sampler, -2.64, and set all $\{1/\xi_{\mu_l}\}_{l=1}^{L}$ equal to half the range of the observed data for all l (a rough estimate of dispersion), 189.6. Sensitivity was assessed by halving and doubling the values of ξ_{μ_l}. Under a similar argument, a_ϕ and b_ϕ should be chosen so that $E\left(1/\phi_l\right) = b_\phi/\left(a_\phi - 1\right)$ is also around half the range of the observations, so we choose $a_\phi = 1.25,\ b_\phi = 47.5$. Note that in every scenario we have employed proper priors, as weakly informative proper priors lead to improved performance and improper priors can lead to paradoxical behavior in mixture models, similar to the well known Bartlett-Lindley paradox in Bayesian model selection.

Next, we consider the prior structure on the weights $\omega_l\left(\mathbf{x}\right)$. As discussed above, the use of a continuation ratio Probit model along with normal priors for the trans-

formed weights is convenient, as it greatly simplifies implementation of the model. In particular, the transformed mixture weights $\{\gamma_l(\mathbf{x})\}$ can be sampled in Session 3.2.3 above from conditionally normal distributions. Hyper-parameter choice is also simplified. A common assumption of basic mixture models for iid data is that all components have the same probability a priori. In the current context in which mixture weights are predictor dependent, a similar constraint can be imposed on the baseline conditional distribution by setting $E(\theta_l^1) = \Phi^{-1}(1/(L-l+1))$. Since we build a hierarchy above heterogeneity parameters to allow information borrowing, the variance of θ^i (for $i = 1, 2, \cdots, 7$), is controlled by the distribution of hyper parameters ϵ^i. In order to make sure the continuation ratio $\Phi(\gamma_l(\mathbf{x}))$ is between 0.001 and 0.998 with 0.99 probability, we would expect $\mathrm{Var}(\theta^i) \approx 1$. Smaller values for $V(\theta^i)$ lead to strong restrictions on the set of weights, discouraging small ones (especially for the first few components in the mixture). On the other hand, larger variances can adversely affect model selection. For the hyper parameter ϵ^i of θ^i, in order to make sure $\mathrm{Var}(\theta^i) \approx 1$, we let $c_i = 6,\ d_i = 5$ so that $\mathrm{E}(1/\epsilon^i) = 1$. Still the prior is very weakly informative, corresponding to a prior sample size of 6 data points. Compared to 20 airlines, 1336 routes and many replicates, the prior gives some stability and very small restrictions.

3.5 Implementation

The data were analyzed using the models described in Subsection 3.1. Fifty mixture components were judged sufficient to flexibility characterize changes in the density across predictors, while limiting the risk of over-fitting. Inferences were robust in our sensitivity analysis for L ranging between 40 and 70, but the quality of the fit, as assessed through the plots described in Section 3 and Section 5, was compromised for $L < 40$.

The Gibbs sampler was run for 100,000 iterations following a 70,000 iteration

burn-in period. Code was implemented in Matlab, and the longest running time was 118h on a 2.96-GHz Intel Xeon E5-2690 computer with 32 cores. This run time could be dramatically reduced by improving efficiency of the code and relying on recent developments in scalable Bayesian computation, but we preferred to use standard Gibbs sampling instead of new and less well established computational methods. Examination of diagnostic plots showed adequate mixing and no evidence of lack of convergence.

3.6 Model Fitting Assessment

We use three methods to assess model fitting: cross validation, posterior predictive checking and visual inspection.

Cross validation is widely used to check the out-of-sample predictive capability of the model and also limits problems like overfitting. Specifically, we use a 3-fold cross validation based on predictive logarithm likelihood, a strict proper scoring for density forecast (see Gneiting and Raftery 2007). The model with highest predictive log-likelihood is shown in Equation 3.2

$$\begin{aligned} \gamma_l(\mathbf{x}) &= \theta_l^1 + \theta_a^2 + \theta_r^3 + \theta_{(a,r)}^4 + \theta_m^5 + \theta_{leg}^6 + f_1\left(dev_{start} \mid \boldsymbol{\theta}^8\right) \\ &\quad + f_2\left(dur \mid \boldsymbol{\theta}^9\right) + f_3\left(\log\left(wgt\right) \mid \boldsymbol{\theta}^{10}\right) \end{aligned} \tag{3.2}$$

where predictor, $log(pcs)$ and airline-leg interactions, are dropped. All the following analyses is based on the estimation of model 3.2. See Appendix for the table of models we checked.

Then, we compare this model with OLS, which is widely used in the previous research of flight delays. First we replicate two data sets predicted by OLS and our model separately, then compare them with the original data. The basic idea of posterior predictive checking is that if the model specification is appropriate, we would expect to see something similar to the real data(Rubin, 1984). From

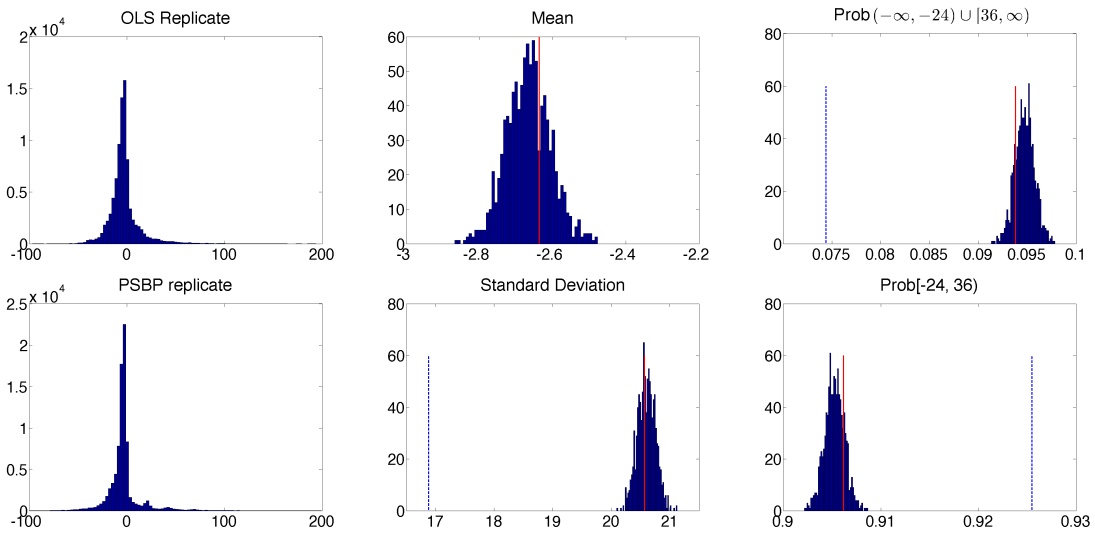

FIGURE 3.2: Posterior predictive model checking

the first column in figure 3.6, the replicated data by OLS resembles the shape of real data poorly: (1) it fails to capture the multimodal features; (2) there is only one single peak much wider also lower than the original data. In sum, the unique extremely narrow while tall major peak, and curvy long-tail features in the original data are all lost in OLS estimation. On the other hand, the replicated data by using our model resembles the real data very well. One step further, we define several test statistics to test our model compared against real data, and also compare to OLS prediction. Specifically, we define test statistics: (1) $\mathsf{E}(y \mid \Psi)$, (2) $\mathsf{Std}(y \mid \Psi)$, (3) $\mathsf{Prob}(y < -24 \text{ or } y \geqslant 36 \mid \Psi)$, (4) $\mathsf{Prob}(-24 \leqslant y < 36 \mid \Psi)$ as shown in Figure 3.6 (blue histograms are PSBP posterior samplers, red solid line is test statistics calculated from real data, purple dashed line is predicted by OLS). Obviously, OLS largely underestimates extreme situations like more than 24 hours earliness or more than 36 hours delays, while overestimating recurrent risk such as deviations between -24 to 36 hours. OLS concentrates at mean estimation, whose prediction is almost the same with data mean (the red line overlays on purple dashed line), while the standard deviation is greatly underestimated. The posterior predictive statistics by

PSBP are close to the true values.

We further check model fitting at a more granular level – airline-route level. In Figure 3.1, the histogram are drawn from real data, blue line is the predictive conditional density by PSBP, while the dotted red line are predicted by OLS. PSBP captures the location and weights of peaks accurately while OLS predicts badly.

4

Results

In Table 4 are the posterior mean and 95% probability interval of (selected) model parameters. There are several things to note from the table:

1. The 50 kernel means, $\mu_1, \mu_2, \cdots, \mu_{50}$, range from -70.0 to 77.5 (hours), indicating the model predicted deviation concentrates within -3 to 3 days, consistent with the data. The 50 kernel standard deviations, $1/\sqrt{\phi_1}, 1/\sqrt{\phi_2}, \cdots, 1/\sqrt{\phi_{50}}$, range from 0.62 to 84.4, meaning the Gaussian kernels can be very narrow or flat, allowing for flexible estimation.

2. Level parameters, $\theta_1^1, \theta_2^1, \cdots, \theta_{49}^2$, vary from -10.9 to 6.74, and the wide range suggests strong variation in risk. For example, if an airline-route pair has $\gamma_l(\mathbf{x}) - \theta_l^1$ close to zero, then for certain l with θ_l^1 smaller than -5, the weight $\propto \Phi(\gamma_l(\mathbf{x})) \approx \Phi(-5) \approx 0$, thus eliminating the inclusion of this component. By similar arguments, θ_l^1 can also help determine, for which $\gamma_l(\mathbf{x}) - \theta_l^1$, component l plays major role.

3. The posterior distributions of coefficients all present substantial learning from their prior distribution, in addition, the 95% probability intervals are narrow.

Table 4.1: Posterior summaries of model parameters

Kernel Parameters				
μ_l $(l = 1, 2, \cdots, 50)$	$\min(\mu_l) = -79.6,$	$\max(\mu_l) = 76.01$		
$1/\sqrt{\phi_l}$ $(l = 1, 2, \cdots, 50)$	$\min(1/\sqrt{\phi_l}) = 0.72,$	$\max(1/\sqrt{\phi_l}) = 84.4$		
Parameters in Weight γ				
Category Predictors				
θ_l^1 $(l = 1, 2, \cdots, 49)$	$\min(\theta_l^1) = -10.9,$	$\max(\theta_l^1) = 6.74$		
θ_a^2 $(a = 1, 2, \cdots, 20)$				
θ_1^2 A1	θ_2^2 A2	θ_3^2 A3	θ_4^2 A4	θ_5^2 A5
0	0.03	-5.27	5.15	3.09
(0, 0)	(-0.40, 0.61)	(-5.86, -4.83)	(4.31, 6.11)	(2.89, 3.26)
θ_6^2 A6	θ_7^2 A7	θ_8^2 A8	θ_9^2 A9	θ_{10}^2 A10
1.16	8.53	2.54	-0.82	2.90
(0.84, 1.53)	(8.19, 8.91)	(2.01, 2.98)	(-1.22, -0.40)	(2.23, 3.64)
θ_{11}^2 A11	θ_{12}^2 A12	θ_{13}^2 A13	θ_{14}^2 A14	θ_{15}^2 A15
-3.35	5.74	-2.96	2.74	-2.82
(-4.02, -2.74)	(5.44, 5.97)	(-3.19, -2.67)	(2.27, 2.98)	(-3.26, -2.36)
θ_{16}^2 A16	θ_{17}^2 A17	θ_{18}^2 A18	θ_{19}^2 A19	θ_{20}^2 A20
4.95	-3.16	-5.36	6.23	-2.34
(4.35, 5.50)	(-3.50, -2.76)	(-6.59, -4.41)	(5.79, 6.67)	(-2.58, -2.12)
θ_{leg}^5 $(leg = 2, 3)$				
θ_2^5	θ_3^5			
-0.29	-0.34			
(-0.38, -0.21)	(-0.47, -0.21)			
Hyper-parameters				
$1/\sqrt{\epsilon^1}$	$1/\sqrt{\epsilon^2}$	$1/\sqrt{\epsilon^3}$	$1/\sqrt{\epsilon^4}$	$1/\sqrt{\epsilon^5}$
4.86	3.39	6.26	7.02	0.64
(3.98, 5.93)	(2.62, 4.44)	(5.86, 6.63)	(6.46, 7.60)	(0.46, 0.90)
$1/\sqrt{\epsilon^6}$				
0.74				
(0.51, 1.10)				

4. The posterior estimation of airline coefficient[1], θ_a^2, shows great heterogeneity, and the large standard deviation, $1/\sqrt{\epsilon^2}$, which measures the variations among airlines, confirm this from one other aspect. Closer inspection reveals that except A1, whose coefficient is fixed at zero for identification, 18 of the remaining 19 airlines' 95% probability intervals don't include 0. Furthermore, many of them are far from zero, implying large impact on risk. However, based on OLS, only 2 of the 19 airlines are significantly different from 0 at 5% confidence level. This huge difference underlies the principle of the two estimation methods. OLS focuses in estimating the effects of predictors on distribution ***mean***, and its results indicate airlines don't necessarily affect the mean of transport risk much. However, PSBP's results show that airlines are playing an important role on selecting and weighting possible kernels, which affects the tail shape, number of peaks, probability of extreme observation etc. These results and comparison once again show that OLS, which cannot detect the airlines' (and some other predictors' including routes' etc) impact on transport risk in this case, would lose considerable valuable information. In order to understand a complex data thoroughly, more sophisticated models, such as PSBP, should be used.

5. Since the number of routes and their interactions with airline are large, 1336 and 587 respectively, we don't include their posterior summaries in Table 4. However, posterior summaries of hyper-parameters standard deviation, $1/\sqrt{\epsilon^3}$, illustrated the large heterogeneity between routes. More importantly, the large standard deviation, $1/\sqrt{\epsilon^4}$, represents possibly huge differences in terms of the distribution of transport risks on the same route while by different airlines. This suggests that a careful selection of carriers can result in dramatically different shipping experiences.

[1] We disguise the names of airlines for confidential reasons. The airline index used here is randomly assigned.

5

Applications

Estimates of predictive conditional probability density functions (Cpdf) is key to generating data-driven operations strategies. In the section, we provide several examples for how posterior Cpdf can aid decision making. We note that the usage of Probit stick-breaking posterior estimation is not restricted to the applications listed here.

5.1 Service Comparison for One Shipment

The most straightforward use of PSBP posterior estimation is to provide predictive conditional distribution of transport risk to shippers based on their predetermined cargo-related variables and selectable service-related variables in Table 2.1. This not only helps the shipper to find a preferable service but also helps the forwarder to set targeting price quote. Assume a customer comes with predetermined cargo-related information $c = \{r, m, wgt\}$ and is choosing from services $s = (a, leg, dur) \in S(c)$, where $S(c)$ is the set of services available given c. Here, even though the initial deviation, dev_{start}, is one of the service-related variables, we set it to 0 because this

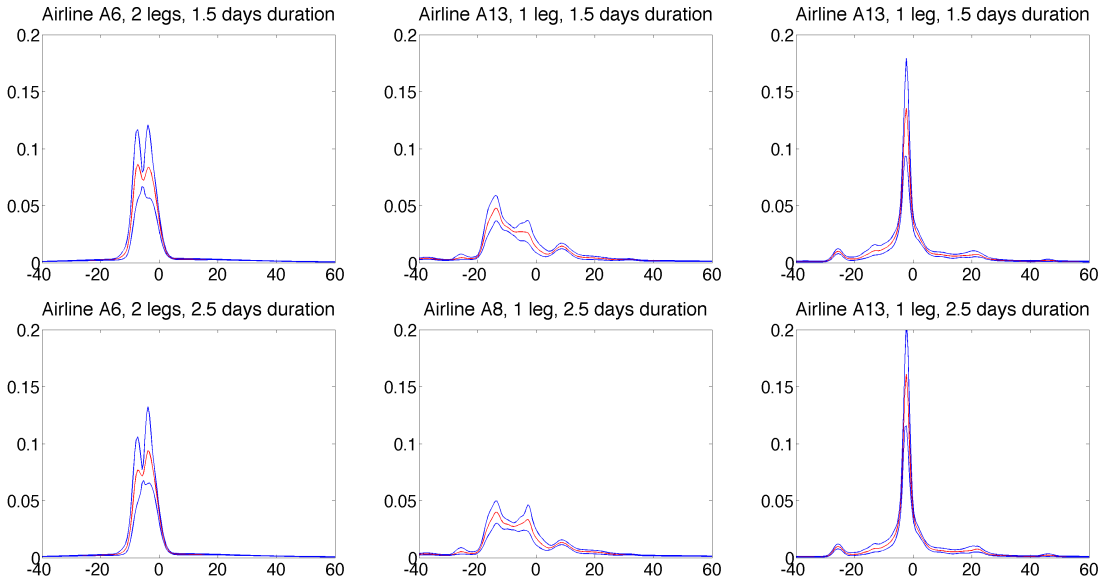

FIGURE 5.1: From Frankfurt (Germany) to Atlanta (United States)

variable is unknown and not selectable before shipping starts. Let $f\left(risk \mid c, s\right)$ be the predictive distribution of transport risk conditional on c and a chosen s, and $l_i\left(risk\right)$ be customer i's loss function. The optimal conditional choice of s, which minimizes transport risks, is defined as

$$\begin{aligned} \left(s \mid c\right)_i^* &\triangleq \operatorname{argmin}_{s \in S(c)} Loss_i\left(s \mid c\right) \\ Loss_i\left(s \mid c\right) &= \int l_i\left(risk\right) f\left(risk \mid c, s\right) \mathrm{d}dev \end{aligned} \tag{5.1}$$

where $Loss_i\left(s \mid c\right)$ is customer i's expected loss of choosing s given c. Estimating each customer's unknown loss function $l_i\left(dev\right)$ is another interesting study of practical value, but is outside the scope of this paper. Here we use several generic loss functions to illustrate how to use predicted $f\left(risk \mid c, s\right)$ to aid service selection.

In Figure 5.1 are the 6 choices as shown by the figure titles, on route from Frankfurt to Atlanta. The choices are randomly picked from the data. We use the following three loss functions

$$l_1\left(risk\right) = C_1 \cdot risk \qquad l_2\left(risk\right) = C_2 \cdot \mathbf{1}\left\{risk > 18\right\} \qquad l_3(risk) = C_3 \cdot risk^2$$

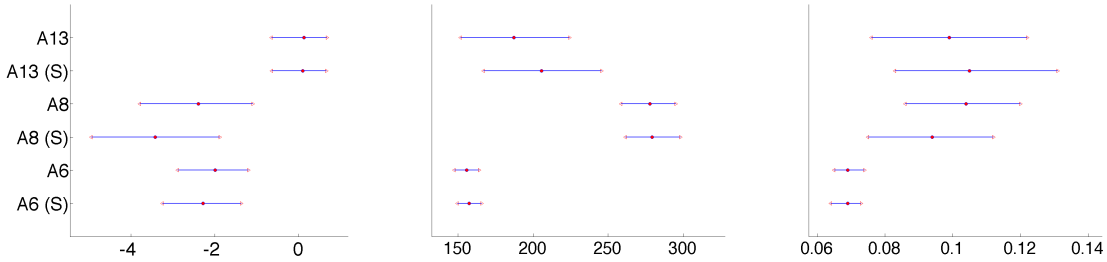

FIGURE 5.2: Ranking based on expected transport risk

l_1 naturally arises when a risk neutral shipper is adverse to delays while fond of early arrivals; l_2 is more proper when a shipper is sensitive to extreme delays exceeding certain threshold (18 hours here); $l3$ is used when a shipper is risk adverse and dislikes any deviations from the plan, neither negative nor positive. Under these loss functions, the expected losses have simple analytical forms

$$Loss_1 = c\mathsf{E}_f \qquad Loss_2 = c\left(1 - F\left(18\right)\right) \qquad Loss_3 = c\left(\mathsf{Var}_f + \mathsf{E}_f^2\right)$$

where f is short for $f\left(risk \mid c, s\right)$ and F is the corresponding cumulative density function. Figure 5.1 presents the expected losses (with 95% probability intervals) calculated for the six choices under 3 risk functions, in which we use (S) to indicate *speedy* service. We observe (1) the rank of services in terms of expected losses varies by loss functions; (2) choice of airlines is playing a more dominant role than the choice between normal and speedy services given an airline.

With estimated expected loss of each choice, forwarders can offer different price quotes to different types of shippers. In this example, a forwarder can lower A8's prices to attract price-sensitive shipper , while increase A6's prices to attract quality-sensitive shippers under loss function 2, thus a higher revenue.

5.2 Supplier Ranking on Route or Higher Level

Unlike a shipper, whose decision is made at the level of each shipment, a forwarder plans its business at the route or higher level. To help solving problems at high

levels, the full conditional predictive Cpdf should be integrated. Specifically, let the full information set be $U = \{a,\ r,\ m,\ leg,\ dur,\ dev_{start},\ wgt\}$, for $U = U_1 \cup U_2$ and $U_1 \cap U_2 = \phi$, then

$$f\left(dev \mid U_1\right) = \int f\left(y \mid U_1, U_2\right) f\left(U_2\right) dU_2$$

$f\left(dev \mid U_1\right)$ is useful when the variables in U_1 are either decision variables or conditional variables. For example, a practical problem faced by a forwarder is whether to choose a carrier on a certain route and how much capacity to reserve from it. For such decisions, an estimation of the airline's service reliability is a critical input. In this case $U_1 = \{a, r\}$ and $U_2 = U - U_1$. By using Equation 5.1 with c and s replaced by r and a, the forwarder can obtain expected losses by each airline $a \in S\left(r\right)$, which, in turn, can help make the right capacity reservation and pricing decisions.

5.3 Baseline Comparison

Our result can also be used to generate baseline comparisons of various factors. Baseline effect of a certain factor excludes the effects of any other factors, thus allowing for direct comparison between factors in one type.

One interesting example is to understand the baseline performance of each airline, in which case a direct comparison is impossible due the fact that airlines serve different routes. To achieve this baseline comparison, we use the average value for all other predictors, except airline effects θ_a^2, as their reference levels, plug them in the posterior samples of each airline and then obtain the reference risk distribution for each airline (See Figure 5.3 for 6 samples from the 20 airlines. See Appendix for the rest 14 baseline distributions). From the plots we can directly compare airlines, which differ from each other by the number of peaks, location of peaks as well as the height of peaks. As such, our model allows baseline comparison based on distribution knowledge. This offers a much richer comparison than those appearing in the

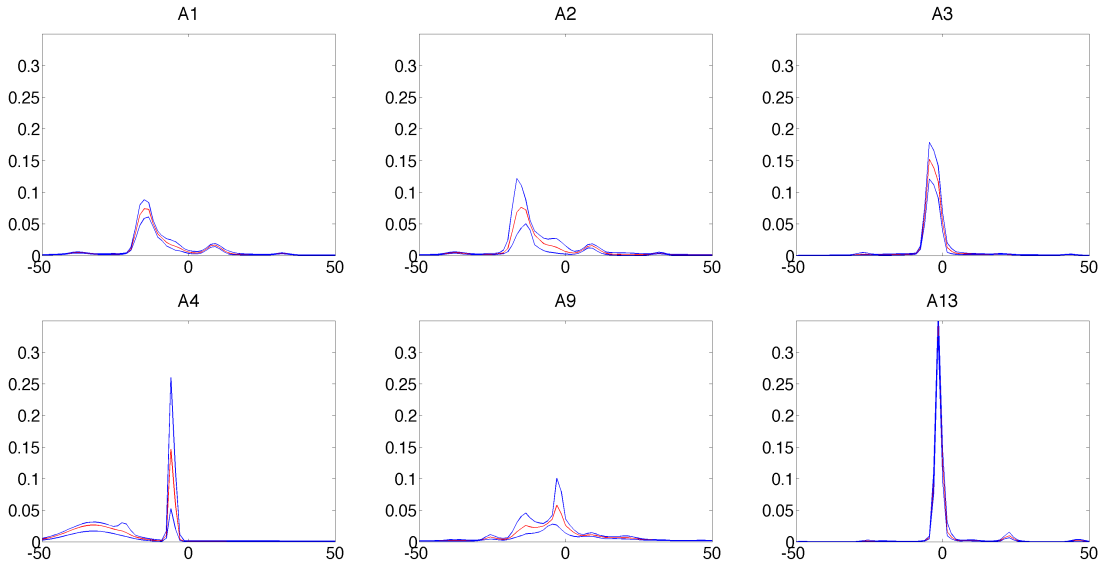

FIGURE 5.3: Sample Airline reference performances

literature based on single average metrics. Meanwhile, the richer tool allows us to obtain the simple metric comparisons as special cases.

For example, using a U.S. passenger flight data set, Deshpande and Arikan (2012) analyze single-leg flight truncated block time, which is transport risk plus planned duration minus initial deviation. Initial deviation is defined as the positive delay of the previous flight by the same craft if applicable and zero otherwise. The authors argue that if the truncated block time is shorter than the scheduled block time, the airline incurs an overage cost of C_o per unit overage time. Otherwise, the airline incurs an underage cost C_u per unit shortage time. The authors then estimate the overage to underage ratio, $\varphi = C_0/C_u$, for each flights, and calculate the mean ratio of flights served by a certain airline as the airline-wise overage to underage ratio, φ_a. Using our international air cargo data, we can obtain an analogous metric by replacing "schedule block time" and "truncated block time" in their paper with *dur* and $(dur + \text{arrival deviation} - [dev_{start}]^+)$. One concern of estimating airline-wise ratio φ_a by simply calculating the average of flight-wise ratios is that the effects from other factors, such as routes etc, cannot be excluded. Thus, the calculated

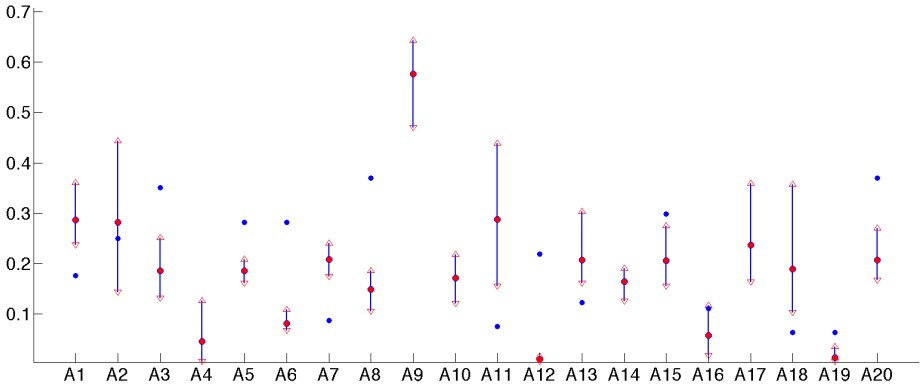

FIGURE 5.4: Overage to underage ratio of airlines

overage to underage ratio of each airline, φ_a, cannot be used for direct comparison of airlines' intrinsic service quality. Baseline distribution of airlines, on the other hand, is a good solution to this problem. Specifically, the optimal dur^* is defined by news-vendor solution that

$$\begin{aligned} \mathsf{Prob}\left(dur^* + \text{arrival deviation} - [dev_{start}]^+ \leqslant dur^* \mid a\right) &= \frac{1}{1+\varphi_a} \\ \mathsf{Prob}\left(\text{arrival deviation} \leqslant 0 \mid a\right) &= \frac{1}{1+\varphi_a} \end{aligned} \tag{5.2}$$

where we use the fact that the reference level of $[dev_{start}]^+$, calculated by the data average, is zero. Thus each airline's overage to underage ratio is calculated by $\varphi_a = \frac{1}{F_a(0)} - 1$; see Figure 5.3 for the calculated overage to underage ratios of 20 airlines with 95% probability intervals.

The overage to underage ratio φ_a is related to airline's on-time probability by Equation 5.2: the higher the on-time rate the lower the ratio φ_a. We compare our results to C2k Monthly Statement issued by IATA. In particular, we choose monthly report issued in November 2012, the same period of our data, and convert the reported airlines' on-time rate into their overage/underage ratio (represented by the blue dots in Figure 5.3). The blue dots deviate from our estimations, the red

dots following no obvious rules. We believe this is because IATA calculated the on-time rate by simply averaging on-time times of an airline, which fails to exclude the impacts from factors other than airline, i.e. cargo weight, route etc, and thus results in unfair comparison. The baseline distribution we calculated can also be used to calculate many other metrics, such as variance, probability of extreme disruptions etc, rather than the simple on-time rate reported by IATA's monthly report.

6

Conclusion and Future Directions

Using data from international air cargo logistics, in this paper, we investigate ways to assess and forecast transport risks, defined as the deviation between actual arrival time and planned arrival time. To accommodate the special multimodal feature of the data, we introduce a Bayesian nonparametric mixture model, the Probit stick-breaking process (PSBP) mixture model, for flexible estimation of conditional density function of transport risk. Specifically, we build a linear structure, including cargo-related variables and service-related variables, into kernel weights so that the probability weights change with predictors. One of the main advantage of the PSBP is its generality, flexibility, relatively simple sampling algorithm compared to other similar ones and consistency under weak regulation conditions. Our results show that this method achieves accurate forecast, while the simpler OLS method can lead to misleading inferences. We also demonstrate how an accurate estimation of transport risk Cpdf can help shippers to choose from multiple available services, and help a forwarder to set targeting price, etc. In addition, we show how to use the model to estimate baseline performance of a predictor, such as airlines' baseline performances. We compare our findings with performance reports issued by IATA

and point out the shortcomings of IATA's simple way of ranking airlines. We note that the usage of our method can be much broader than the examples shown here. Indeed, any decisions involving a distribution function needs an estimated Cpdf.

One of the interesting findings of our paper is that airlines have critical impact on the shape of transport risk distribution rather than the mean focused by OLS. For future research, we hope to be able to obtain data containing information regarding why and how airlines are performing so differently on the same route. By knowing the root drivers of airline service performances we can actually improve the service quality rather than simply choose the best performer.

Appendix A

Data

A.1 Data Cleaning

After matching MUP with its baseline RMP, we obtain 155,780 shipments (matching rate is higher than 95%). After dropping (1) shipments with extremely delayed milestones (usually caused by data input error); (2) shipments missing critical information (e.g. carrier); (3) shipments missing weight or package information[1], 139,512 shipments are retained. The 139,512 shipments are operated by 20 airline on 11,282 routes (and B to A are two distinct routes), and form 17,604 airline-route pair (each airline-route pair means this airline is operating on that route). Since our analysis is conducted on each airline-route pair level, in order to avoid the high noisy caused by sparse observations, we drop route-airline pairs containing less than 10 observations and routes containing less than 20 observations in the observing period (half a year). After applying the filter, we have 86150 observations left operated by 20 airlines on 1,333 routes. The filter is effective in selecting large and profitable route.

[1] refer to Appendix A for more details about data cleaning steps

A.2 Data Illustration

In Table A.1 are the current members under C2K standards. In Table A.2 is a typical

Airlines

- Air Bridge Cargo [+]
- Air Canda (C)
- Air France (C)
- Alitalia [+]
- American (C)
- Austrian Airlines [#]
- Avianca [+]
- British Airways (C)
- Blue 1 [#]
- Cargolux (C)
- Cargolux Italia [#]
- Cathay Pacific (C)
- China Airlines [+]
- China Southern [t]
- Delta (C)
- Dragonair [#]
- Ethiopian Airlines [+]
- Etihad (C)
- Finnair (C)
- Garuda Indonesia [+]
- Iberia (C)
- Kenya Airways [+]
- KLM (C)
- Korean (C)
- LACSA [#]
- Lufthansa (C)
- Martinair [#]
- Polar [+]
- Qantas [+]
- Qatar Airways (C)
- SAS (C)
- Saudia [+]
- Singapore (C)
- South African Airways [+]
- South African Airways [+]
- Swiss (C)
- TACA International/ Peru [#]
- Tampa Cargo [#]
- Turkish Airlines (+)
- United (C)
- Virgin Atlantic [t]

Forwarders

- Agility Logistics (C)
- Aramex [t]
- Cargomind [t]
- CEVA [t]
- DHL Global Forwarding (DD)
- Hellmann [+]
- Kuehne + Nagel
- JAS Worldwide [+]
- Panalpina (C)
- Schenker AG (DD)
- SDV Intl. Logistics (C)
- Uti (Spain) [+]
- Yusen Air & Sea

FIGURE A.1: Cargo 2000 members

Table A.1: An example of a route map

Milestone	Time	Airport	Flight	Weight	Piece
RCS	06.12.2013 16:15:00	NTE	#	630	2
DEP	06.12.2013 19:00:00	NTE	AA 8854	630	2
ARR	07.12.2013 08:52:00	CDG	AA 8854	630	2
DEP	10.12.2013 09:21:00	CDG	AA 0063	630	2
RCF	10.12.2013 21:26:00	MIA	AA 0063	630	2
DEP	11.12.2013 14:58:00	MIA	AA 0913	630	2
RCF	11.12.2013 21:46:00	BOG	AA 0913	630	2
DLV	11.12.2013 22:40:00	BOG	#	630	2

route map for a shipment from Nantes (France) to Bogot (Columbia). In Figure A.2 is the milestone chain and explanation for each milestone. In Table A.2 is an typical

Table A.2: A typical record of exception

Status	Exception	Time	Flight	Airport
DEP	COCSYMD	08.01.2013 05:05:00	BA 0125	LHR

record of an exception.

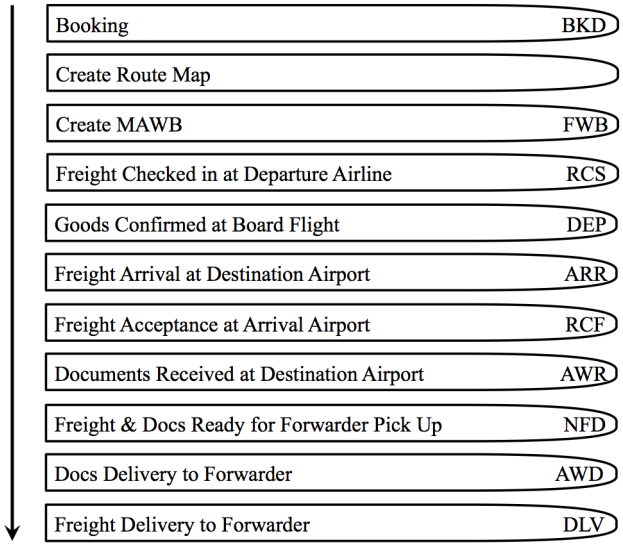

FIGURE A.2: Milestone explanations

Appendix B

Supporting Algorithm and Material

B.1 Label Switching

1. From $1, 2, \ldots, L$ choose two elements l_1 and l_2 uniformly at random and change their labels with probability

$$\min\left(1, \Pi_{\mathbf{x}\in\mathcal{X}}\left(\frac{\omega_{l_1}(\mathbf{x})}{\omega_{l_2}(\mathbf{x})}\right)^{n_{l_2}(\mathbf{x})-n_{l_1}(\mathbf{x})}\right)$$

where $n_l(\mathbf{x}) = \sum_j s_j(\mathbf{x}) = l \ (j = 1, \cdots, n(\mathbf{x}))$

2. Sample a label l uniformly from $1, 2, \ldots, L-1$ and propose to swap the labels l, $l+1$ and corresponding stick-breaking weights γ_l, γ_{l+1} with probability

$$\min\left(1, F \times \Pi_{\mathbf{x}\in\mathcal{X}}\frac{\left(1-\Phi\left(\gamma_{l+1}(\mathbf{x})\right)\right)^{n_l(\mathbf{x})}}{\left(1-\Phi\left(\gamma_l(\mathbf{x})\right)\right)^{n_{l+1}(\mathbf{x})}}\right)$$

where

$$F = \frac{\mathsf{N}\left(\theta_l^1 \mid \Phi^{-1}\left(\frac{1}{L-l}\right), 1\right) \cdot \mathsf{N}\left(\theta_{l+1}^1 \mid \Phi^{-1}\left(\frac{1}{L-l+1}\right), 1\right)}{\mathsf{N}\left(\theta_l^1 \mid \Phi^{-1}\left(\frac{1}{L-l+1}\right), 1\right) \cdot \mathsf{N}\left(\theta_{l+1}^1 \mid \Phi^{-1}\left(\frac{1}{L-l}\right), 1\right)}$$

is the change of prior probability since the prior of θ^1 is not symmetric.

Table B.1: Cross validation for model comparison

	Model	-2LL		Model	-2LL
1	Ξ	18807	6	$\Xi - \theta^7_{(a,leg)} - \theta^4_{(a,r)}$	18949
2	$\Xi - \theta^7_{(a,leg)}$	18452	7	$\Xi - \theta^7_{(a,leg)} - \theta^6_{leg} - \theta^{11}$	18533
3	$\Xi - \theta^7_{(a,leg)} - \theta^6_{leg}$	18576	8	$\Xi - \theta^7_{(a,leg)} - \theta^5_m - \theta^{11}$	18480
4	$\Xi - \theta^7_{(a,leg)} - \boldsymbol{\theta}^{11}$	18439	9	$\Xi - \theta^7_{(a,leg)} - \theta^6_{leg} - \theta^5_m$	18976
5	$\Xi - \theta^7_{(a,leg)} - \theta^5_m$	18497	10	$\Xi - \theta^7_{(a,leg)} - \theta^4_{(a,r)} - \theta^2_a$	19067

B.2 Label Switching for Finite Mixture Model

1. Sample a label l uniformly from $1, 2, \ldots, L-1$ and propose to swap the labels l, $l+1$ and corresponding stick-breaking weights γ_l, γ_{l+1} with probability

$$\min\left(1, F \times \Pi_{\mathbf{x}\in\mathcal{X}} \frac{\left(1-\Phi\left(\gamma_{l+1}\left(\mathbf{x}\right)\right)\right)^{n_l(\mathbf{x})}}{\left(1-\Phi\left(\gamma_l\left(\mathbf{x}\right)\right)\right)^{n_{l+1}(\mathbf{x})}}\right), \text{ if } l \leqslant L-2$$

where

$$F = \frac{f\left(\alpha_l \mid \Phi^{-1}\left(\frac{1}{L-l}\right), 1\right) f\left(\alpha_{l+1} \mid \Phi^{-1}\left(\frac{1}{L-l+1}\right), 1\right)}{f\left(\alpha_l \mid \Phi^{-1}\left(\frac{1}{L-l+1}\right), 1\right) f\left(\alpha_{l+1} \mid \Phi^{-1}\left(\frac{1}{L-l}\right), 1\right)}$$

is the change of prior probability and $f(\cdot \mid \mu, \phi)$ is the probability density function of $\mathsf{N}(\cdot \mid \mu, \phi)$. If $l = L-1$, the Metropolis-Hasting probability is:

$$\min\left(1, \Pi_{\mathbf{x}\in\mathcal{X}}\left[\frac{\Phi\left(\gamma_l\left(\mathbf{x}\right)\right)}{1-\Phi\left(\gamma_l\left(\mathbf{x}\right)\right)}\right]^{n_{l+1}(\mathbf{x})-n_l(\mathbf{x})}\right), \text{ if } l = L-1$$

B.3 Cross Validation

In Table B.3 is the cross validation results calculated for each model by using 10,000 samples with the first 10,000 samples dropped as burn-in, where we use Ξ to indicate the full model in Equation 3.1 and use "$-$" to indicate dropping certain predictors. We use LL to indicated average predictive log-likelihood. Specifically, based on 3-fold cross validation, for each model, we calculate the predictive log-likelihood of the left-out data for three times, and use the average of these three log-likelihoods as

the LL of this model. Since we are comparing $-2LL$, so the smaller the value the stronger the predictive capability of that model. Thus, we choose model (4) in the Table B.3.

B.4 Supporting Figures

In Figure B.4 are the baseline risk distributions of the rest 14 airlines.

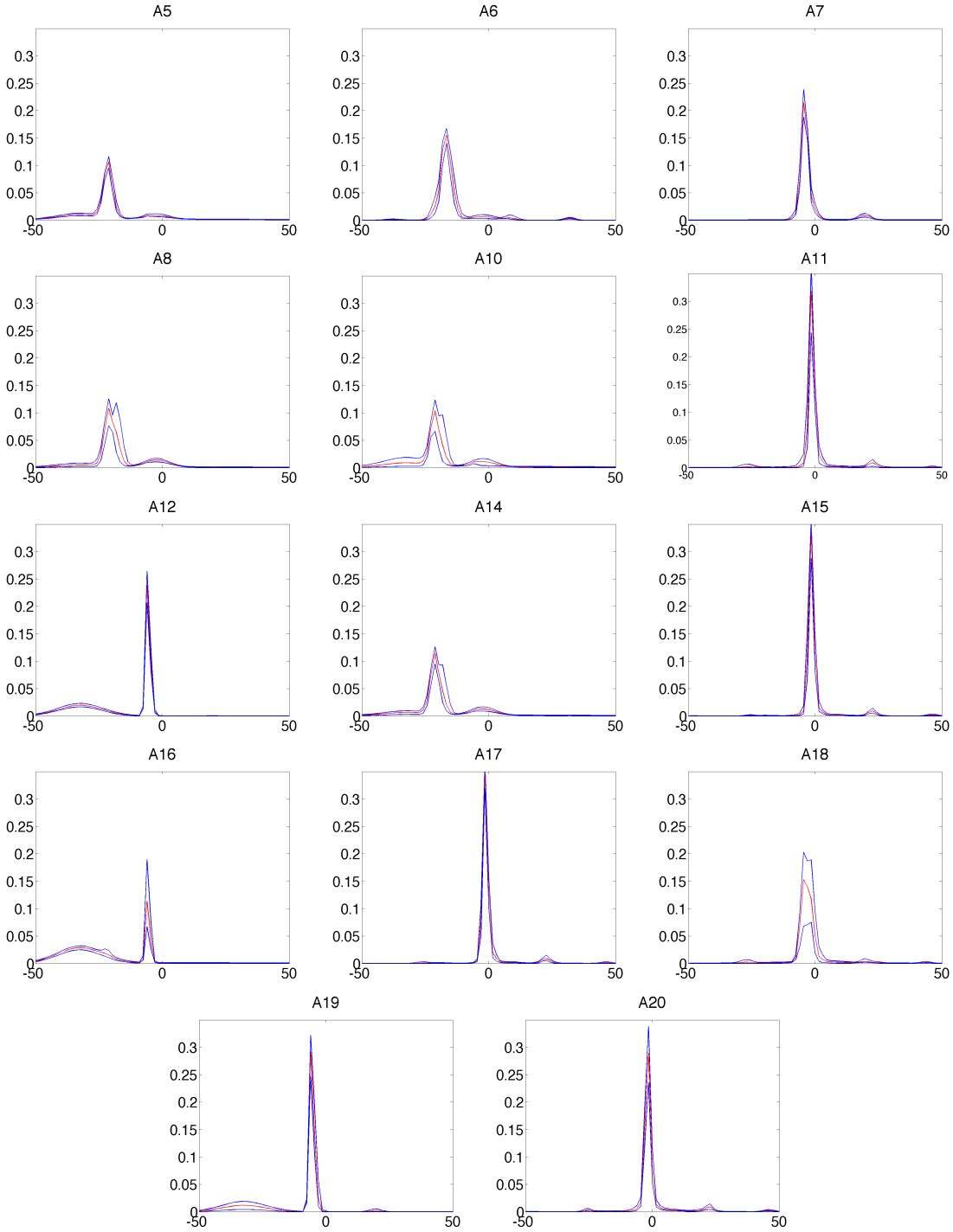

FIGURE B.1: Airline reference performances

Bibliography

Chung, Y. and Dunson, D. B. (2009), "Nonparametric Bayes Conditional Distribution Modeling with Variable Selection," *Journal of the American Statistical Association*, 104.

Cohen, M. A. and Kunreuther, H. (2007), "Operations risk management: overview of Paul Kleindorfer's contributions," *Production and Operations Management*, 16, 525–541.

Crabtree, T., Edgar, J., Hoang, T., Tom, R., and Hart, B. (2012), "World Air Cargo Forecast 2012-2013," Industry report, The Boeing World Air Cargo Forecast Team.

Deshpande, V. and Arikan, M. (2012), "The Impact of Airline Flight Schedules on Flight Delays," *Manufacturing & Service Operations Management*, 14, 423–440.

Escobar, M. D. and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Fan, J., Yao, Q., and Tong, H. (1996), "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems," *Biometrika*, 83, 189–206.

Federgruen, A. and Yang, N. (2009), "Optimal supply diversification under general supply risks," *Operations Research*, 57, 1451–1468.

Gneiting, T. and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378.

Guajardo, J. A., Cohen, M. A., Kim, S.-H., and Netessine, S. (2012), "Impact of Performance-Based Contracting on Product Reliability: An Empirical Analysis," *Management Science*, 58, 961–979.

Gupta, D. (2008), "Flexible carrier - forwarder contracts for air cargo business," *Journal of Revenue & Pricing Management*, 7, 341–356.

Hall, P., Wolff, R. C. L., and Yao, Q. (1999), "Methods for Estimating a Conditional Distribution Function," *Journal of the American Statistical Association*, 94, 154–163.

Hyndman, R. J. and Yao, Q. (2002), "Nonparametric Estimation and Symmetry Tests for Conditional Density Functions," *Journal of Nonparametric Statistics*, 14, 259–278.

IATA (2014), "C2K Master Operating Plan," .

IATA (2014), "IATA - Cargo," .

Ishwaran, H. and Zarepour, M. (2002), "Dirichlet Prior Sieves in Finite Normal Mixtures," *Statistica Sinica*, 12, 941–963.

Kleindorfer, P. R. and Saad, G. H. (2005), "Managing disruption risks in supply chains," *Production and operations management*, 14, 53–68.

Kleindorfer, P. R., Belke, J. C., Elliott, M. R., Lee, K., Lowe, R. A., and Feldman, H. I. (2003), "Accident Epidemiology and the US Chemical Industry: Accident History and Worst-Case Data from RMP* Info," *Risk Analysis*, 23, 865–881.

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates," *The Annals of Statistics*, 12, 351–357.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. (2011), "Big data: the next frontier for innovation, competition, and productivity," Tech. rep., McKinsey Global Institute.

Morrell, P. S. (2011), *Moving boxes by air: the economics of international air cargo*, Ashgate Publishing Limited, Farnham Surrey England.

Mueller, E. R. and Chatterji, G. B. (2002), "Analysis of aircraft arrival and departure delay characteristics," in *AIAA aircraft technology, integration and operations (ATIO) conference*.

Muller, P., Erkanli, A., and West, M. (1996), "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, 83, 67–79.

Papaspiliopoulos, O. (2008), "A note on posterior sampling from Dirichlet mixture models," .

Papaspiliopoulos, O. and Roberts, G. O. (2008), "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models," *Biometrika*, 95, 169–186.

Pati, D., Dunson, D. B., and Tokdar, S. T. (2013), "Posterior consistency in conditional distribution estimation," *Journal of Multivariate Analysis*, 116, 456–472.

Rodriguez, A. and Dunson, D. B. (2011), "Nonparametric Bayesian models through probit stick-breaking processes," *Bayesian Analysis*, 6, 145–177.

Rodriguez, A., Dunson, D. B., and Taylor, J. (2009), "Bayesian hierarchically weighted finite mixture models for samples of distributions," *Biostatistics*, 10, 155–171.

Rubin, D. B. (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician," *The Annals of Statistics*, 12, 1151–1172.

Shumsky, R. A. (1995), "Dynamic statistical models for the prediction of aircraft take-off times," Thesis, Massachusetts Institute of Technology, Thesis (Ph. D.)–Massachusetts Institute of Technology, Sloan School of Management, 1995.

Tomlin, B. (2006), "On the Value of Mitigation and Contingency Strategies for Managing Supply Chain Disruption Risks," *Management Science*, 52, 639–657.

Van Mieghem, J. A. (2009), "Risk Management and Operational Hedging: An Overview," *The Handbook of Integrated Risk Management in Global Supply Chains*, pp. 13–49.

Wang, Y., Gilland, W., and Tomlin, B. (2010), "Mitigating supply risk: Dual sourcing or process improvement?" *Manufacturing & Service Operations Management*, 12, 489–510.